

průběh práce s českými

≈ s h a i yš. □
| X s T ÷ T k □ D n



Day 1
Image modality

B 1/2 Y E A
B 1/2 1/2 E S T S C A d T
E E U E E



Day 5
3D models



B 1/2 Y E
° a S || || n 1/2 U Y



B 1/2 Y E
L d T || || n 1/2 U Y

≈ S L ' 1 / 2 ° I I U S L I J C S I B '

ê j ʒ | | ñ 1 / 2 ʒ | | ñ s ʒ ʒ 1 / 2 ʒ | | s i l ʒ ʒ e e s s n ʒ ʒ e e | | 1 / 2 ʒ | | ŷ ° s e i t e i j i i ŷ ° s e p i 1 / 2 e s n | | ʒ | | s e i j e a | | ʒ e
| | ʒ | | j i e | | ñ 1 / 2 ʒ | | s e i

' S B S S ° e j e a | | ʒ

ñ s s ° ʒ ʒ e e e e | | j ʒ | | ñ 1 / 2 ʒ | | j i e e
ü a | | ʒ | | s e i ʒ 1 / 2 ʒ | | ŷ s e

2. Early Fusion

multimodal inputs are fed to the model rather than to its internals

' S s c i 1 / 2 ʒ | | ñ 1 / 2 ʒ | | ; | | e e s
i i s ʒ | | ʒ | | e 3 B c P

' S s e j e i | | ʒ
ü ʒ ŷ s e i j ü B c P e

2.1. Non-tokenized (NT-EF)

ʒ s e | | ʒ | | s e i ʒ | | e 3 L c P e

û ° s ʒ ʒ 1 / 2 ʒ | | ʒ | |

- ° s c s s s c e 1 / 2 ʒ | | ° ʒ c
- L c | | e e s ʒ | | i i s ʒ | | ʒ | |
- i 1 / 2 ʒ | | e 1 / 2 ʒ | |



ê | | L 3 ü 1 / 2 ʒ | |

- ü e | | ʒ | | ñ s c e ü
- ê | | L e ʒ ŷ s c
- ≈ | | j i s c



f ü s ʒ ʒ ° ü

- A e i 1 / 2 s e c s i c ʒ ʒ ʒ | |
- s ʒ | | ñ s c | | e e e | | ʒ e
| | ñ 1 / 2 ʒ | | ŷ

ü 1 / 2 p e

- i | | ʒ | | s e i ʒ | | e e
ü e i 1 / 2 ʒ | | ñ 1 / 2 ʒ | | ŷ
- i | | ʒ | | e s s l i | | e e
| | s e s c



ÜSLĪĴ ĆŠĒ ½

1 e½ĒĴĒĒ½ĴĴ ŠĴĴ½ĴĴ ½ĴĴĒ

2 cĴĴ pĴ ½ŠĒ ° ĳŠĴ 3Ü½ pĒ

Ā SOTA models for video modality

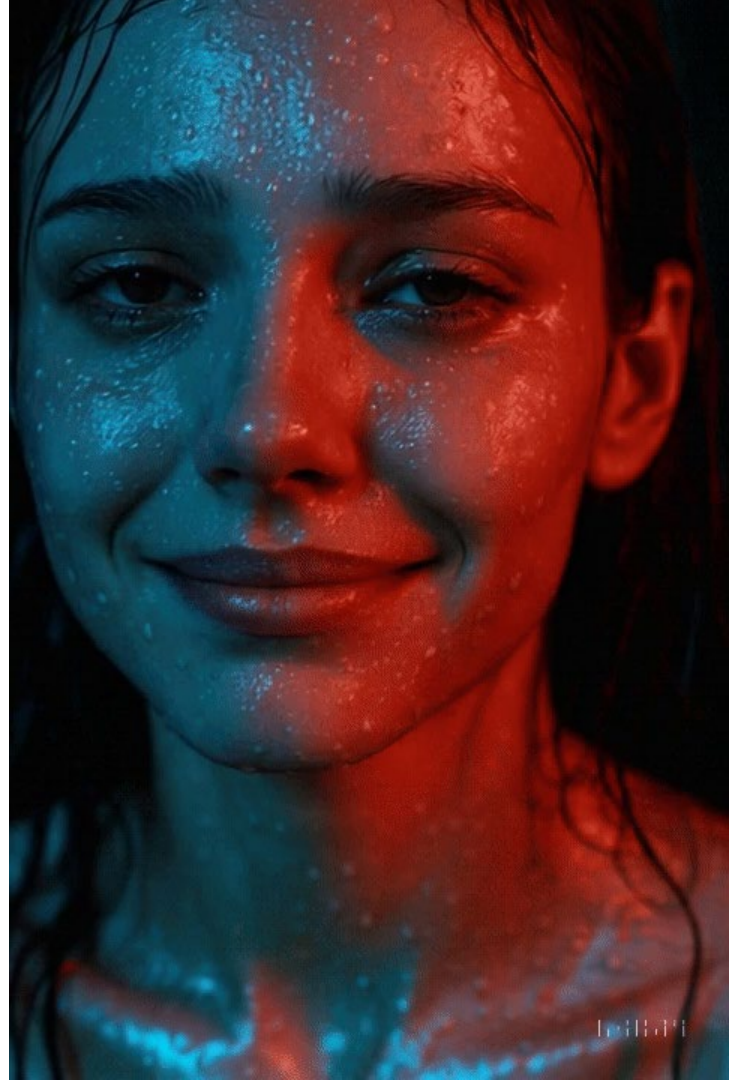
E ĴŸĴĴ 3½ĴŸĴĴ Ĵ ½ĴŸ



·

: $\hat{S} = \mathbb{L} \uparrow \downarrow \mathbb{L} \parallel \frac{1}{2} \hat{G} \parallel \dot{E}$

Tasks and benchmarks to
test models with video
modality





: S=L↑|| ½G||EIE ã S|| 3ê ê LHIT, “EP

pT JIEF|| G|| ½d|| T̄ ùãS|| IE|| J̄ T̄ IEJ í Ĩã S̄EIE|| IE|| J̄ T̄ Ē|| ñ ½YP



Video-MME

On what date did the individual in the video leave a place that Simon thought was very important to him?

- A. May 31, 2022.
- B. June 9, 2021.
- C. May 9, 2021.
- D. June 31, 2021.

The date of **Day 1** is May 31, 2021.
[in Frames]



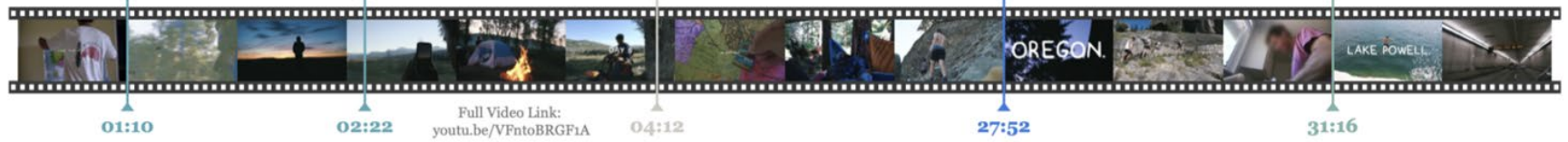
Simon is the camera man.
[in Frames]



Yosemite National Park did mean a lot more to Simon. [in Subs/Audio]

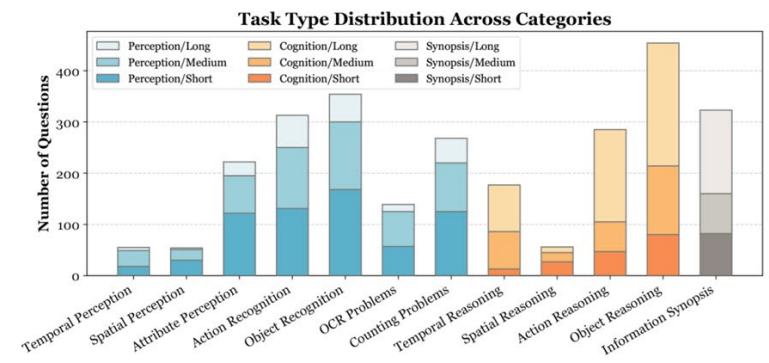
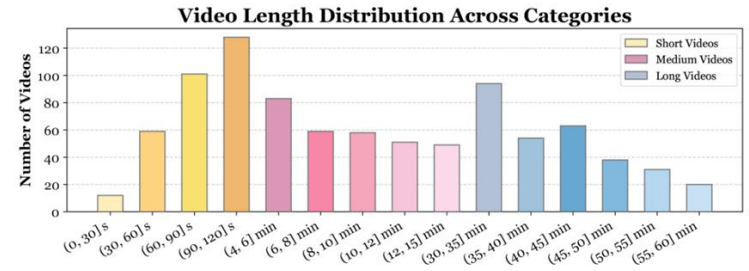
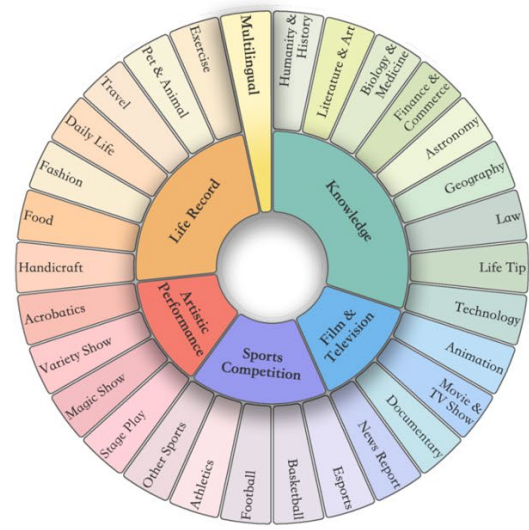


Depart Yosemite on **Day 10**.
[in Frames]



: Š=ŦL↕|| ½ĜŦEİĒ ã Š|| 3ê ê LĤŦ, “EP

- BĴĈĀĬŦK
 !Ĭ|| ĒĒ|| ̄ŦĴİSĒĒ
 ĨĒĒĴĴĈ
- c ΔĤSĈŋ ĩ ½Ĭ½Ē
 !Ĭ|| ĨĴĵeĴİ ŠİĒ
 ŠūSĈŋ ½ŸĒŠİĒ
 € ĩĒSĈSĒ



: S=LL↑|| 1/2G||EIEE 𐌹 S|| 3e ê LHIT, "EP

–E|| 𐌹 E3'HE|| 𐌹 30-60 min

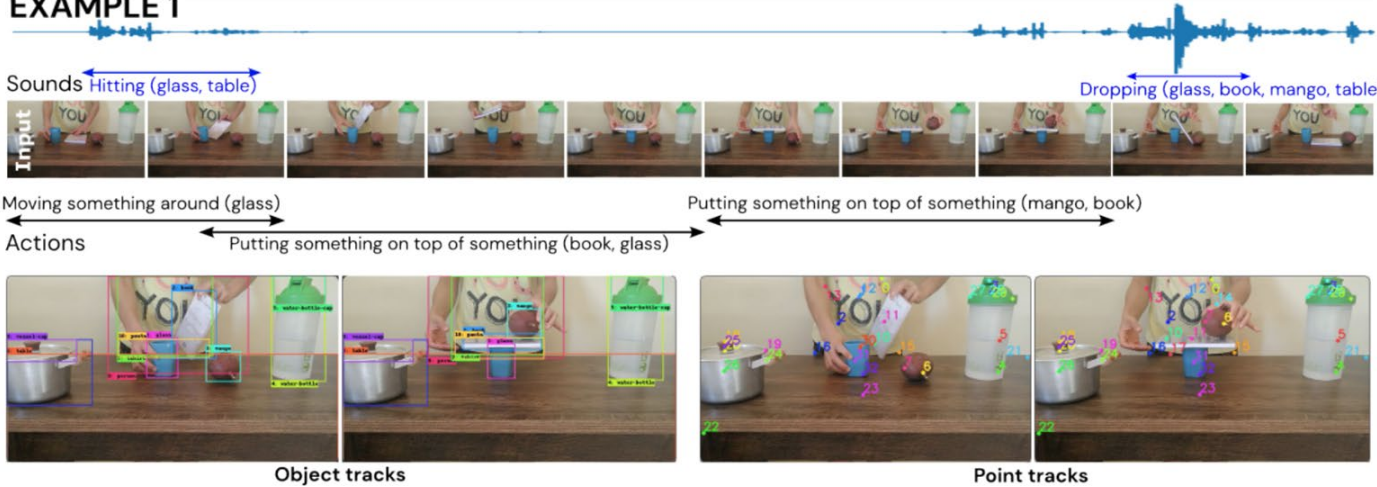
#	Model	LLM Params	Frames	Date	Overall (%)		Short Video (%)		Medium Video (%)		Long Video (%)	
					w/o subs	w subs	w/o subs	w subs	w/o subs	w subs	w/o subs	w subs
1	Gemini 1.5 Pro Google	-	1/0.5 fps ^{1*}	2024-06-15	75.0	81.3	81.7	84.5	74.3	81.0	67.4	77.4
2	AdaReTaKe HIT & Huawei	72B	1024	2025-03-04	73.5	79.6	80.6	82.8	74.9	79.7	65.0	76.4
3	Qwen2-VL Alibaba	72B	768 ^{3*}	2024-08-19	71.2	77.8	80.1	82.2	71.3	76.8	62.2	74.3
4	GPT-4o OpenAI	-	384 ^{2*}	2024-06-15	71.9	77.2	80.0	82.8	70.3	76.6	65.3	72.1
5	LLaVA-Video Bytedance & NTU S-Lab	72B	64	2024-08-28	70.6	76.9	81.4	82.8	68.9	75.6	61.5	72.5

- authors update the benchmark manually
- we will analyze open-source models
- benchmark does not necessarily require strong long-context skills



: $\hat{S} = \mathbb{L} \uparrow \parallel \frac{1}{2} \hat{G} \uparrow \text{EIE} \hat{S} \hat{C} \hat{L} \hat{S} \circ \hat{I} \hat{d} \hat{=} \hat{\epsilon} \hat{S} \hat{E} \hat{I} \hat{H} \hat{I} \text{ , } \text{“ AP}$

EXAMPLE 1



- $\circ \hat{m} \hat{S} \hat{I} \hat{r} \hat{1} \hat{2} \hat{J} \hat{n} \hat{d} \hat{E}$
 $\hat{I} \hat{C} \hat{L} \hat{r} \hat{E} \hat{H} \hat{I} \hat{I} \hat{=} \hat{w} \hat{S} \hat{E} \hat{I} \hat{E}$
 $\hat{n} \hat{I} \hat{E} \hat{P}$
- $\hat{e} \hat{J} \hat{L} \hat{o} \hat{S} \hat{E}$
 $\hat{L} \hat{r} \hat{=} \hat{d} \hat{S} \hat{=} \hat{1} \hat{E} \hat{E} \hat{V} \hat{E}$
 $\hat{n} \hat{S} \hat{I} \hat{S} \hat{C} \hat{I} \hat{=} \hat{r} \hat{E} \hat{I} \hat{d} \hat{E}$
 $\hat{=} \hat{S} \hat{I} \hat{C} \hat{E}$

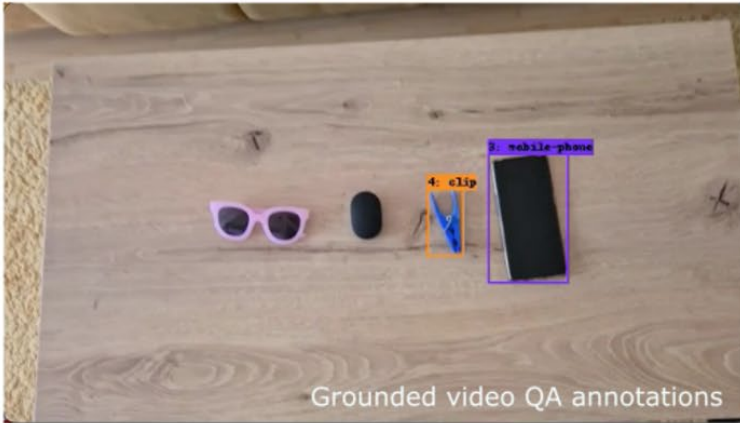
Multiple-choice video QA
Area: Physics, Reasoning: Predictive
Question: Is the configuration of objects likely to be stable after placing the last object?

Options:
 a) The configuration is likely to be stable.
 b) The configuration is likely to be unstable.
 c) One cannot judge the stability of this configuration.



: $\hat{S} = \mathbb{L} \uparrow \parallel \frac{1}{2} \hat{G} \uparrow \text{EiE} \rightarrow \hat{S} \hat{C} \hat{L} \hat{S} \circ \tilde{I} \mathbb{L} \parallel \mathbb{E} \hat{S} \hat{E} \tilde{I} \hat{H} \hat{I} , \sim \hat{A} \hat{P}$

$\mathbb{E} \hat{S} \hat{E} \hat{I} \hat{E} \hat{E} \parallel \hat{S} \parallel \mathbb{L} \hat{C} \hat{Y} \parallel ! \hat{H} \uparrow \hat{S} \hat{E} \parallel \parallel \hat{n} \hat{S} \mathbb{L}$



Multiple-choice video QA

Area: Memory, Reasoning: Explanatory

Question: What changed on the table while the camera was looking away?

Options:

- a) The mobile and clip swapped positions.
- b) The bottle and watch were removed and a clip and mobile were added.
- c) The mobile was added and a clip was removed.

Grounded video QA

Area: Memory, Reasoning: Descriptive

Question: Track the objects that were added to the table while the camera was looking away.

Benchmarks: PerceptionTest (2023)

€SÉĪĒĒ|| Ś|| || ŐŸ|| !Ī↓ŚĒ|| || ņŚ ħĪĪ|| Ĵ=ĪĪ↓ŚĒ|| í śĻĪĒĒ



Multiple-choice video QA
Area: Memory, Reasoning: Counterfactual
Question: *If the person had put the objects in the backpack in reverse order, which object or objects would have been put in second?*
Options: a) shirt b) pen c) laptop



: S=L↑|| 1/2G1EIE ã S 3e ê ê , HÏ , 'HP

Video Lecture

105

1275

395

1865

2935 **[In Video Quiz]**

Can you fill in the superscripts and subscripts for the second neuron?

✓ $a_2^{[3]} = g(w_2^{[3]} \cdot a^{[2]} + b_2^{[3]})$ ← in-video quiz answer

✗ $a_2^{[3]} = g(w_2^{[3]} \cdot a^{[2]} + b_2^{[3]})$ input is $a^{[2]}$

✗ $a_2^{[3]} = g(w_2^{[3]} \cdot a_1^{[2]} + b_2^{[3]})$ input is a vector $a^{[2]}$ not a single number

4655 **[In Video Quiz]**

Activation value of layer l , unit (neuron) j : $a_j^{[l]} = g(w_j^{[l]} \cdot a^{[l-1]} + b_j^{[l]})$

output of layer $l-1$ (previous layer)

general formula

sigmoid activation function

Parameters w & b of layer l , unit j

Quiz Time

Perception

Can you **identify key information** in the video?

➤ The in-video quiz answer is _____. The general formula for the in-video-quiz is _____.

✓ I can **perceive** that the quiz answer is $a_2^{[3]} = g(w_2^{[3]} \cdot a^{[2]} + b_2^{[3]})$. The formula is $a_j^{[l]} = g(w_j^{[l]} \cdot a^{[l-1]} + b_j^{[l]})$.

Comprehension

Can you **comprehend the knowledge** introduced in the video?

➤ Based on your understanding, can you fill in the superscripts and subscripts for the neuron in layer 4?

✓ I can **comprehend** the general formula. By substituting layer=4, the answer is $a_1^{[4]} = g(w_1^{[4]} \cdot a^{[3]} + b_1^{[4]})$.

Adaptation

Can you **adapt what you learned from the video to solve a novel yet related problem**?

➤ <Case 1> is NOT covered in the video, can you calculate the output of neuron 4?

✓ I can **adapt** what I learned from the video to solve <Case 1>. The answer is 0.117.



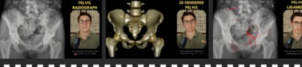




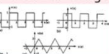
Quiz

<Case 1>

: S=L↑|| 1/2G1EIE ã ð ð ð
 ā J S E I ð ð ð ð ð ð ð ð ð
 È ½ || ◦ S ñ È ð || È
 J ð ð ð ð ð ð ð ð
 İ S w İ || ð È

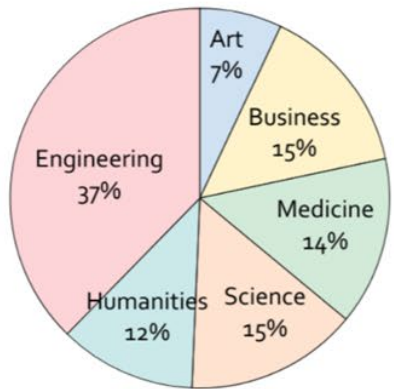
: Š=Ł↑|| ½ĜıEİİĚ ăŃŠ 3ê ê ê , ĨT , ‘HP

- Videos from YouTube
- Choose from 6 topics
- Topics are divided into perception and problem solving
- Multiple choice

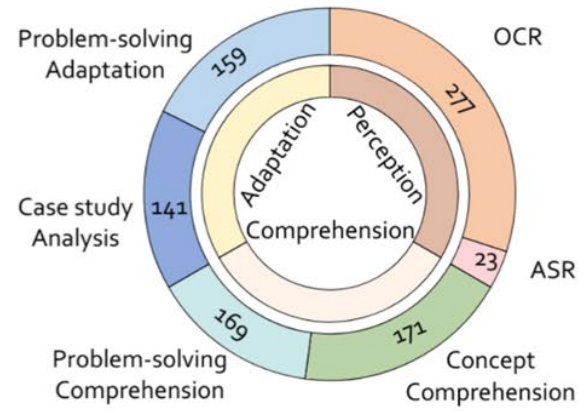
Art	Humanities	Medicine
 <p>Question: What does the speaker say when introducing Peter Paul Rubens at the end of the video? Select the option that precisely matches the speaker's statement. Options: (A) Peter Paul Rubens was a famous Baroque... (B) Peter Paul Rubens is regarded as a prolific artist... (I) Peter Paul Rubens was the most important... (J) Peter Paul Rubens is celebrated for his dynamic...</p> <p>Track: Perception, Video Type: Concept-introduction video, Subject: Art Theory, QA Type: Automatic Speech Recognition (ASR)</p>	 <p>Question: Based on your understanding of cultural universals from the video, determine which of the following statements are correct: Statement 1: All human cultures have some... Statement 2: The video uses the example of... Statement 3: At 3:35, the video implies that ... Statement 4: ... Statement 5: ... Options: (A) Statement 1 (B) Statement 2,3 (C) Statement 3,4 (D) Statement 2,4,5(J) Statement 2,4</p> <p>Track: Comprehension, Video Type: Concept-introduction video, Subject: Sociology, QA Type: Concept Comprehension (CC)</p>	 <p>Question: Can you identify the abnormality on this plain film of the pelvis? <image 1>  Options: (A) Bone cyst (B) Acute hip fracture (C) Osteoarthritis (D) Surgical hardware (E) Resection of the pubic symphysis ... (J) Bone infection</p> <p>Track: Adaptation, Video Type: Concept-introduction video, Subject: Clinical Medicine, QA Type: Case Study Analysis (CSA)</p>
Business	Science	Engineering
 <p>Question: According to the video, a minimum price control on alcoholic drinks is intended to reduce consumption from Q1 to ____, addressing negative externalities. The policy raises the price to ____ above the free market price of _____. Fill in the blanks based on the video content. Options: (A) Q*, Pmin, P1 (B) Q*, P1, Pmin (C) Q1, Pmin, P2 (D) Q2, P1, Pmin (E) Q*, P2, P1 ... (F) Q1, P2, Pmin (G) Q2, Pmin, P1. (H).... (I).... (J) Q1, P1, Pmin</p> <p>Track: Perception, Video Type: Problem-solving video, Subject: Economics, QA Type: Optical Character Recognition (OCR)</p>	 <p>Question: In the video, Example Question (1) is solved with an angle $\theta=25$ degrees. If the angle θ is adjusted to 30 degrees while all other conditions remain unchanged, what will be the updated result for Example Question (1) as explained in the video? Options: (A) 4.00 seconds (B) 2.82 seconds (C) 3.50 seconds (D) 2.50 seconds (E) 3.04 seconds (F) 2.00 seconds (G) 3.15 seconds (H) 1.85 seconds (I) 2.25 seconds (J) 3.85 seconds</p> <p>Track: Comprehension, Video Type: Problem-solving video, Subject: Math, QA Type: Problem-solving Strategy Comprehension (PSC)</p>	 <p>Question: Based on what you learned from the video, write the Fourier series for the three voltage waveforms in (a) of <image 1>.  Options: (A) $(4/\pi)(\sin(\pi t)+(1/2)\sin(3\pi t)+(1/4)\sin(5\pi t)+\dots)$ (B) $(4/\pi)(\sin(\pi t)+(1/3)\sin(3\pi t)+(1/5)\sin(5\pi t)+\dots)$ (C) $(4/\pi)(\sin(\pi t)+(1/2)\sin(2\pi t)+(1/4)\sin(4\pi t)+\dots)$ (J) $(4/\pi)(\sin(\pi t)+(1/4)\sin(3\pi t)+(1/6)\sin(5\pi t)+\dots)$</p> <p>Track: Adaptation, Video Type: Problem-solving video, Subject: Electronics, QA Type: Problem-solving Strategy Adaptation (PSA)</p>

: S=L↑|| ½G¶EiEë ã S|| 3ê ê ê , HïT , “HP

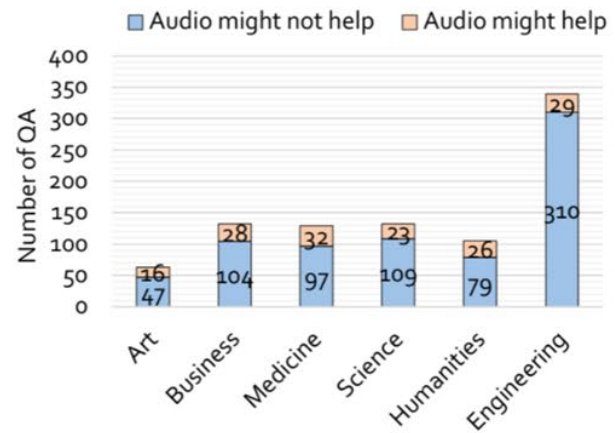
How much the model understood **what happened overall** / questions about the **concept** in the video / **adaptation** of knowledge to a new but similar task



(a) Video distribution across disciplines.



(b) QA distribution across types.



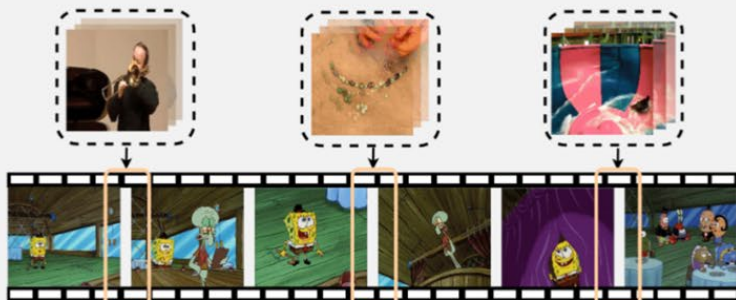
(c) QA distribution with respect to audio.

Benchmarks: MLVU (2025)

n n Ćn 1/4 || Ƴ || ! Ƴ || Ć || 1/4 || Ƴ || Ć || Ƴ n ĵ Ć Ƴ 1/2 Ƴ ĵ 1/4 || Ƴ ||

Multi-Detail LVU

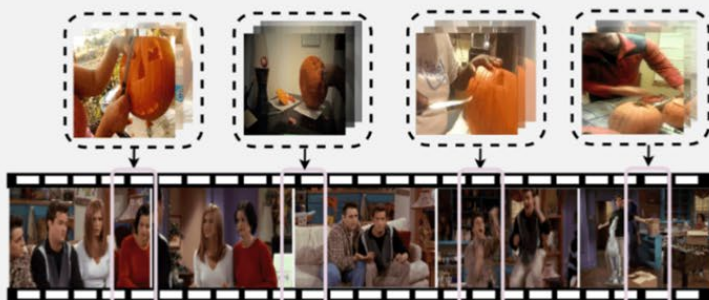
(h) Action Order



Q: Order these actions from the video: (1) water skiing, (2) playing trombone, (3) making jewelry.

- (A) 1 -> 2 -> 3 (B) 1 -> 3 -> 2 (C) 2 -> 1 -> 3 **(D) 2 -> 3 -> 1**

(i) Action Count

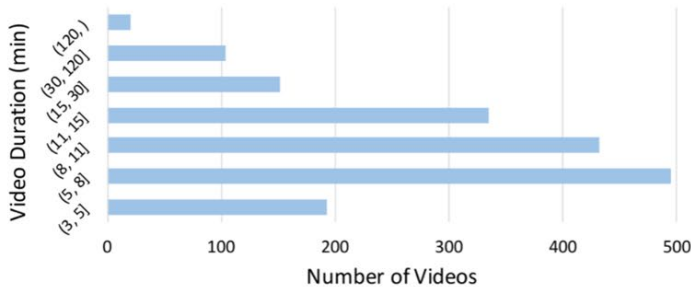


Q: How many times does the action of "carving a pumpkin" occur in this video?

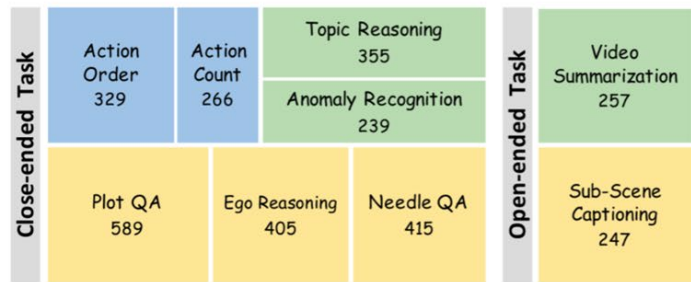
- (A) 0 (B) 2 **(C) 4** (D) 6

: $\hat{S} = \mathbb{L} \updownarrow \parallel \frac{1}{2} \hat{G} \updownarrow \text{EiE} \ddot{U}$, $\ddot{H} \updownarrow$, $\ddot{H} \updownarrow$

≈ s h u y s̄ . □ □
 | X̄ s̄ T̄ ÷ T̄ k̄ □ X̄ n



° \hat{S} | $\ddot{E} \ddot{U} \updownarrow \ddot{E} \ddot{E}$
 $\updownarrow \hat{J} \ddot{C} \ddot{E} \ddot{E} \hat{J} \ddot{H} \ddot{E} \Delta \ddot{H} \updownarrow \ddot{E}$
 $\hat{E} \parallel \frac{1}{2} \hat{Y}$



■ Holistic Task
 ■ Single-Detail Task
 ■ Multi-Detail Task



: $\hat{S} = \mathbb{L} \uparrow \parallel \frac{1}{2} \hat{G} \uparrow \text{EiEe } \ddot{U} , \ddot{H} \ddot{T} , \ddot{H} \ddot{P}$

- Compare with random guessing
- Open generation (reference, embeddings, LLM)

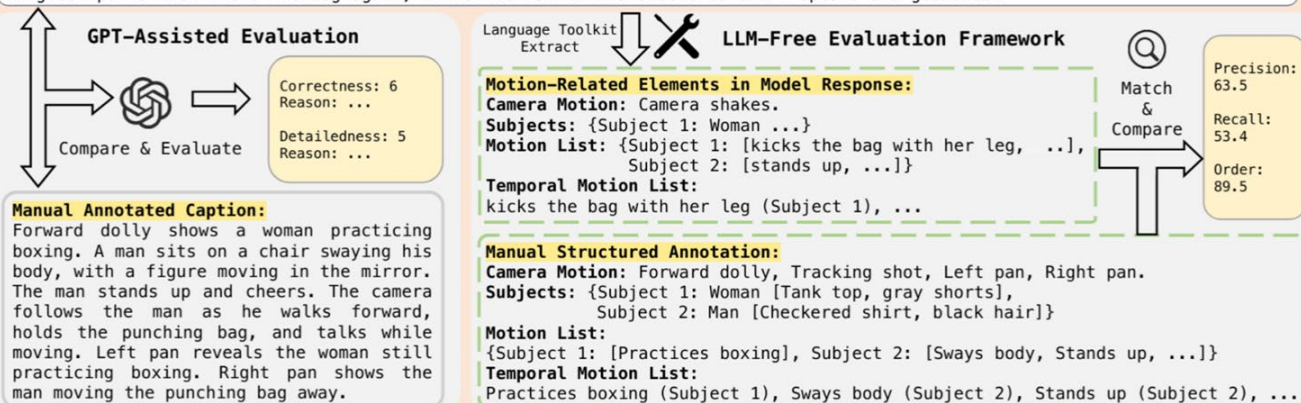
Methods	Date	Input	Holistic			Single Detail				Multi Detail		M-Avg	G-Avg
			TR	AR	VS*	NQA	ER	PQA	SSC*	AO	AC		
Full mark	-	-	100	100	10	100	100	100	10	100	100	100	10
Random	-	-	16.7	16.7	-	16.7	16.7	16.7	-	16.7	16.7	16.7	-
<i>Short Video MLLMs</i>													
Otter-V [23]	2023-05	16 frm	16.5	12.8	2.18	16.7	22.6	22.0	4.20	12.9	13.3	16.7	3.19
mPLUG-Owl-V [54]	2023-04	16 frm	25.3	15.4	2.20	6.7	13.2	22.0	5.01	14.3	20.0	16.7	3.61
VideoChat [25]	2023-05	16 frm	26.4	12.8	2.15	18.3	17.0	22.0	4.90	15.7	11.7	17.7	3.53
Video-LLaMA-2 [59]	2024-08	16 frm	52.7	12.8	2.23	13.3	17.0	12.0	4.87	15.7	8.3	18.8	3.55
VideoChat2-HD [26]	2024-06	16 frm	74.7	43.6	2.83	35.0	34.0	30.0	5.14	21.4	23.3	37.4	3.99
Video-LLaVA [28]	2023-11	8 frm	70.3	38.5	20.9	2.30	26.4	26.0	5.06	20.0	21.7	29.3	3.68
ShareGPT4Video [7]	2024-05	16 frm	73.6	25.6	2.53	31.7	45.3	38.0	4.72	17.1	8.3	34.2	3.63
VideoLLaMA2 [9]	2024-06	16 frm	80.2	53.8	2.80	36.7	54.7	54.0	5.09	42.9	16.7	48.4	3.95
<i>Long Video MLLMs</i>													
MovieChat [41]	2023-07	2048 frm	18.7	10.3	2.30	23.3	15.1	16.0	3.24	17.1	15.0	16.5	2.77
Movie-LLM [42]	2024-03	1 fps	27.5	25.6	2.10	10.0	11.3	16.0	4.93	20.0	21.7	18.9	3.52
LLaMA-VID [27]	2023-11	1 fps	20.9	23.1	2.70	21.7	11.3	16.0	4.15	18.6	15.0	18.1	3.43
MA-LMM [16]	2024-04	1000 frm	44.0	23.1	3.04	13.3	30.2	14.0	4.61	18.6	13.3	22.4	3.83
MiniGPT4-Video [3]	2024-04	90 frm	64.9	46.2	2.50	20.0	30.2	30.0	4.27	15.7	15.0	31.7	3.39
LongVA [60]	2024-06	256 frm	81.3	41.0	2.90	46.7	39.6	46.0	4.92	17.1	23.3	42.1	3.91
Video-CCAM [11]	2024-08	96 frm	79.1	38.5	2.65	45.0	52.8	56.0	4.49	24.3	26.7	46.1	3.57
Video-XL [40]	2024-09	256 frm	78.0	28.2	3.40	50.0	41.5	46.0	5.02	48.6	31.7	46.3	4.21
LLaVA-Onevision [24]	2024-08	32 frm	83.5	56.4	3.75	46.7	58.4	58.0	5.09	35.7	23.3	51.7	4.42
GPT-4o [†] [37]	2024-05	0.5 fps	83.7	68.8	4.94	42.9	47.8	57.1	6.80	46.2	35.0	54.5	5.87

: Š=ŦL↑∥ ½Ĝ∥EİİĊ ° û ≈3: Š=ŦL↑ĤİŦ , “ŦP

Open-Ended Evaluation

Prompt: Please analyze and describe the temporal dynamics in this video, focusing on the camera motions, actions, activities, and interactions rather than the static content.

Model Response: The camera shakes as a woman in green pants hits a punching bag with her hands, then turns to the right and kicks the bag with her leg. A man in a checkered shirt stands up and walks to the bag and holds it steady for her. The woman in green pants then kicks the bag again, while the man in the checkered shirt speaks and gestures.



d -€Ē ĈĒL ĈŦ Ē
İ↓ŠĒ/ŦĒŦĒĒ ĈĒ
◦ Š=ŦŦŦŦŠŦ Ē
n Š=ŦŦĒİŦŦ

2

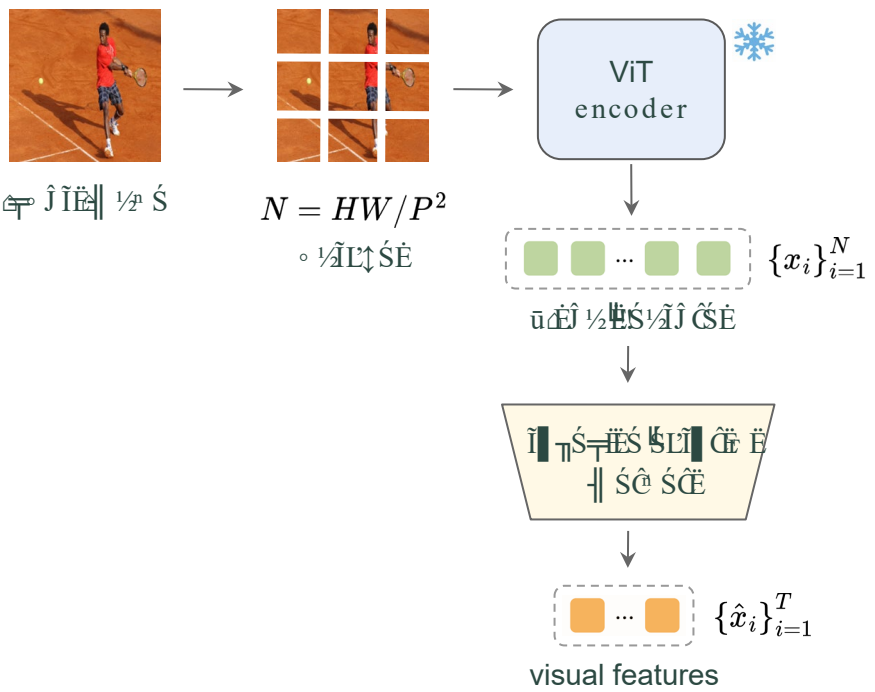
° ě Š ■ 3Ü½° p€

c □ || ě || ½ Š Ě ■ ě ě Š ■ ě
|| ■ ě ½ Ÿ Ÿ Š Ÿ Š = ě = ě ! Ě ↓ Š ě
Ü½° p€ ě || ■ ě Š ě ■ ě ě ě Š ■



ViT: Vision Transformer

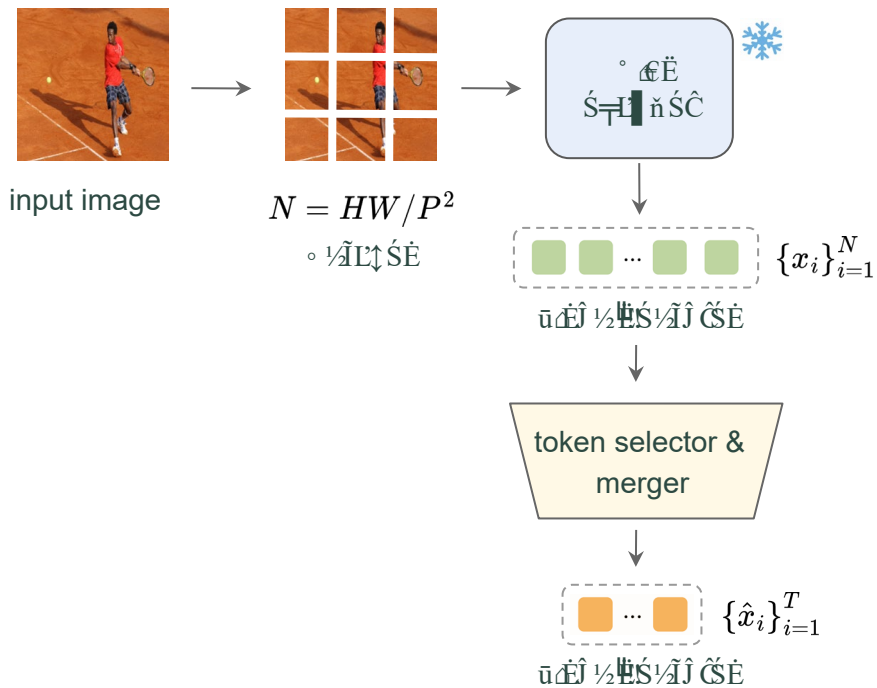
Input image I is divided into patches of size $P \times P$. The total number of patches is $N = HW/P^2$. These patches are processed by a ViT encoder to produce a sequence of tokens $\{x_i\}_{i=1}^N$. These tokens are then processed by a ViT decoder to produce a sequence of visual features $\{\hat{x}_i\}_{i=1}^T$.



Visual Question Answering (VQA)



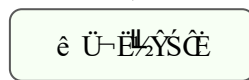
Visual Question Answering (VQA) involves processing an input image and a question to generate an answer.



Visual Question Answering (VQA)



visual features



$\pi \in \mathcal{R}^{N \times 2}$

attention weights

$M \in \{0, 1\}^N$



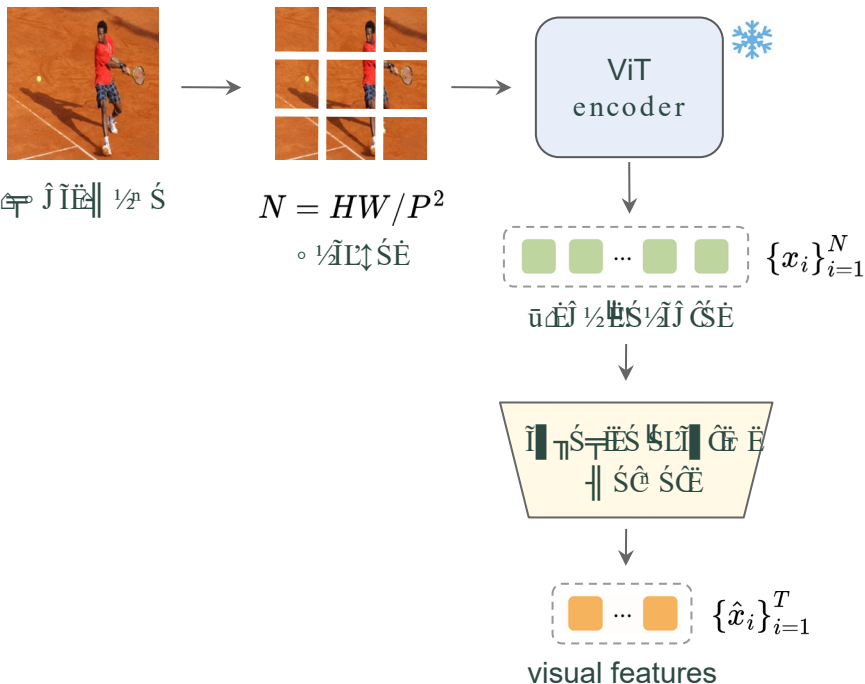
output visual features

ViT: Video Transformer

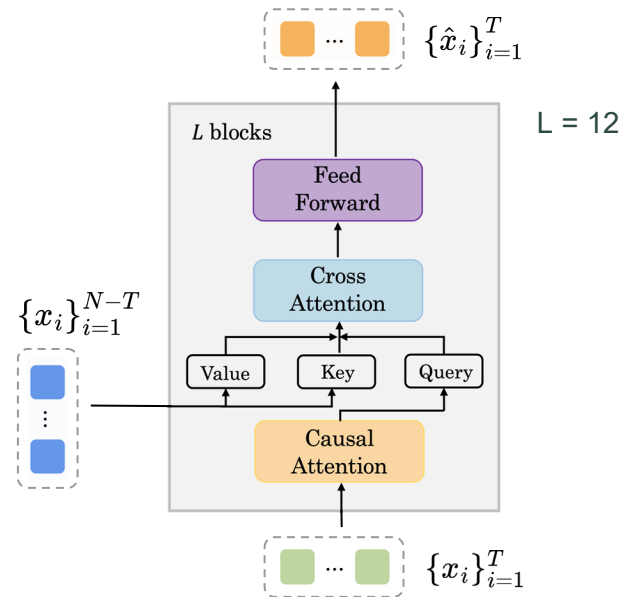
≈ s h a i y s . □
 § l a y b k 1 h a s b l i a b
 ÷ ≈ ≈ ÷ _



u d e j t e j | n s h e d e e 3 d r e i | i l ° 3 ; u p - i i e l / 2 p j 1 / 2 s e j | n s l a i i u 1 / 2 e 3 0



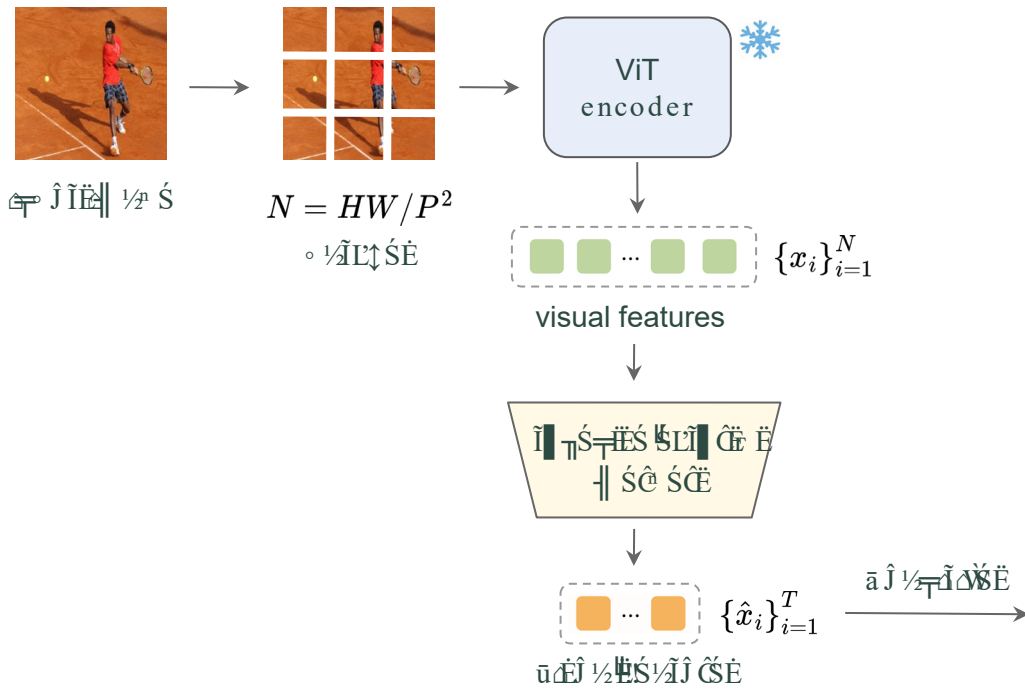
ViT Decoder Architecture



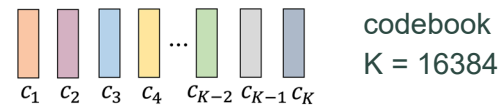
Visual Question Answering with ViT



Visual Question Answering (VQA) with ViT



Visual Question Answering with ViT



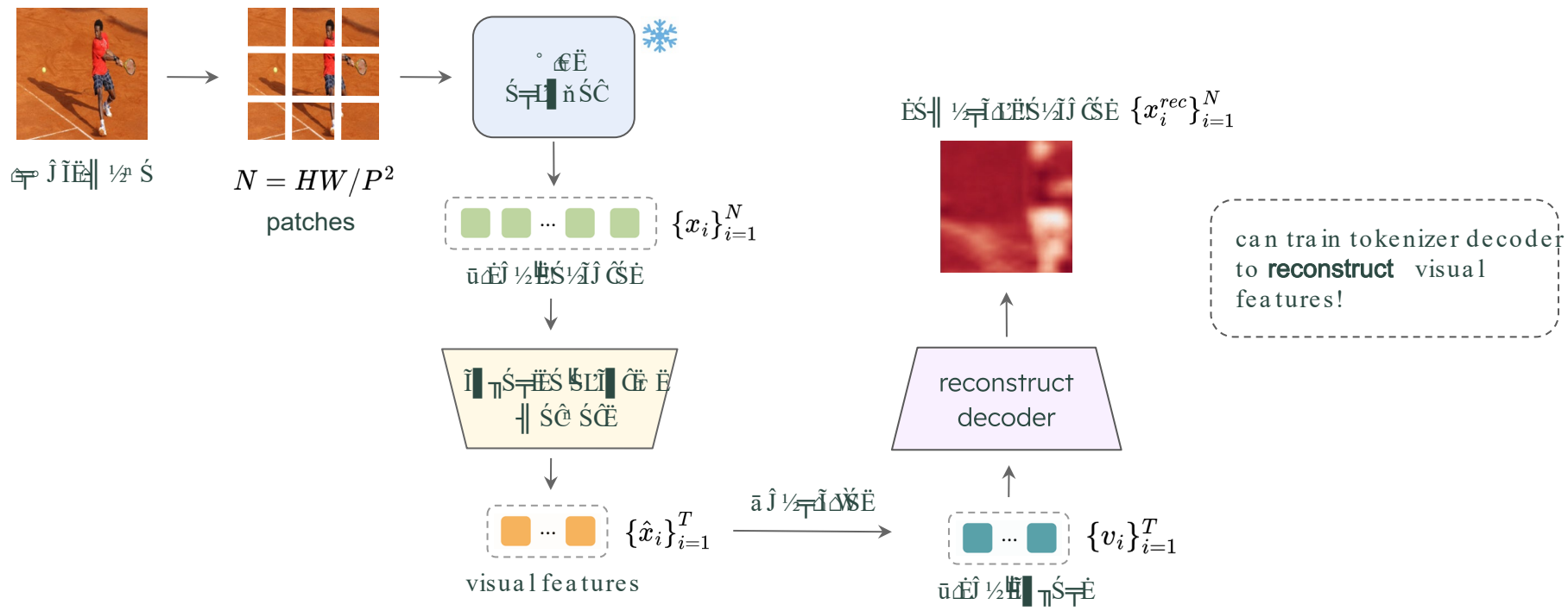
$$v_i = \arg \min_j \|l_2(\hat{x}_i) - l_2(c_j)\|_2$$

$$v_i \in [0, K - 1]$$



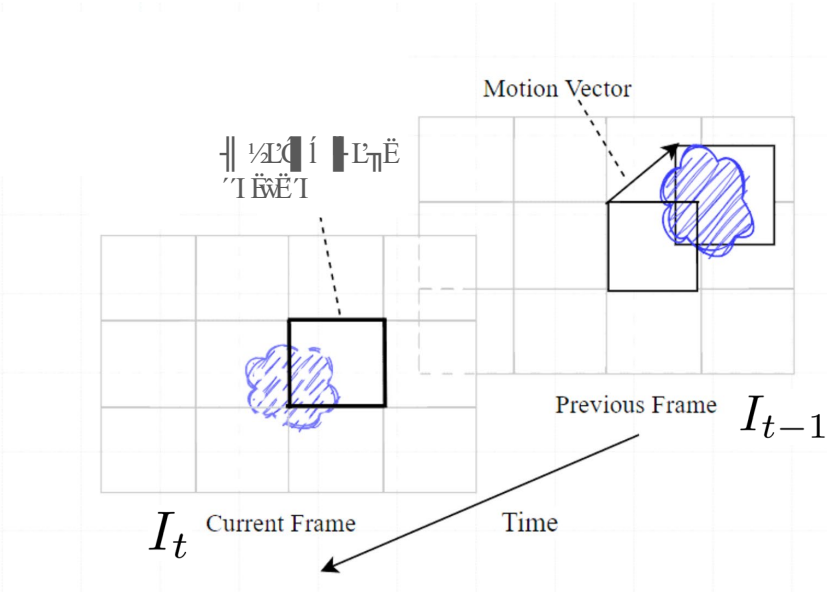
€3Lc İ Ü½° p€İİ ½ĈĒ , “’Eİİp; Ü≈P

ūāĒ ½ĈĒ ĩ ų ĤāĒ €3d Γ’Eİ !İĒ° 3; Üp-İİİ½ĈĒ ĵ ½ ĤĈĒ ĩ ų ĤāĒ İİÜ½ĈĒ 30.



Video -LaVIT: MPEG-4

$L \parallel \circ \parallel \hat{Y} \uparrow \hat{S} \hat{e} - L d 3 E \Pi \Omega \Pi \hat{I} \parallel \hat{h} \hat{m} \hat{m} \hat{S} \hat{H} \uparrow \hat{S} \hat{E} \parallel \frac{1}{2} \hat{S} \hat{H} \parallel \hat{E} \hat{H} \hat{S} \hat{Y} \hat{C} \hat{z} \parallel \hat{S} \hat{E} \Pi \hat{C} \parallel \frac{1}{2} \hat{C} \hat{Y} \hat{E} \hat{S} \parallel \frac{1}{2} \hat{I} \hat{L} \hat{E} \hat{P} \hat{z} \hat{m} \hat{E}$
 $\parallel \parallel \hat{I} \hat{d} \hat{T} \hat{I} \hat{S} \parallel \circ \parallel \hat{C} \hat{z} \hat{S} \hat{u} \parallel \hat{S} \hat{S} \parallel \hat{S} \hat{T} \hat{P}$



$$\vec{m}(p, q) = \arg \min_{i, j} \|I_t(p, q) - I_{t-1}(p - i, q - j)\|$$

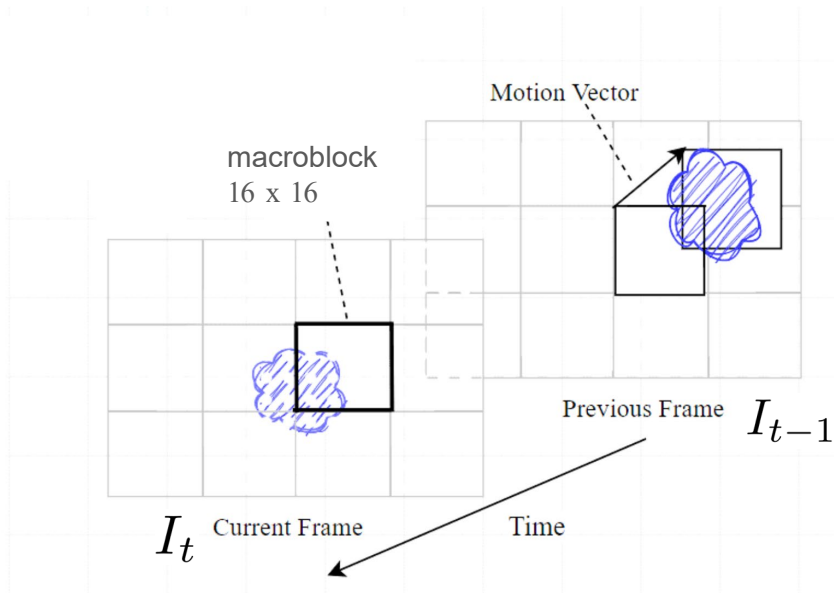
(i, j) — coordinate offset between the center
 $\parallel \hat{E} \hat{E} \parallel \frac{1}{2} \hat{C} \hat{I} \parallel \hat{L} \hat{T} \hat{E}$

$$\hat{u} \hat{m} \hat{S} \parallel \hat{E} \hat{L} \hat{b} \longrightarrow I_0 \in \mathbb{R}^{H \times W \times 3}$$

???

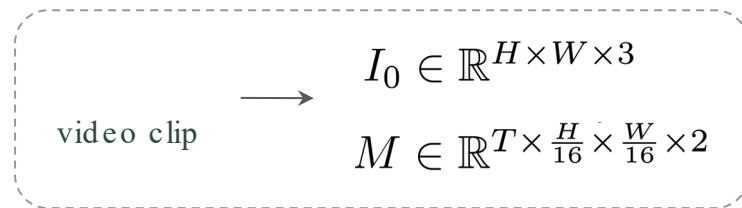
◦ $\hat{m} \hat{S} \parallel 3U \frac{1}{2} \circ p \in \hat{I} \hat{E} \hat{E} \neg L d 3E$

L $\parallel \circ \parallel \hat{Y} \hat{H} \uparrow \hat{S} \hat{e} \neg L d 3E \parallel \Omega \Omega P \hat{I} \parallel \hat{H} \hat{m} \hat{m} \hat{S} \hat{H} \uparrow \hat{S} \hat{E} \parallel \frac{1}{2} \hat{S} \hat{H} \parallel \hat{E} \hat{H} \hat{S} \hat{Y} \hat{C} \hat{z} \parallel \hat{S} \hat{E} \hat{H} \hat{C} \parallel \frac{1}{2} \hat{Y} \hat{E} \hat{S} \parallel \frac{1}{2} \hat{I} \hat{L} \hat{E} \hat{P} \hat{z} \hat{m} \hat{E}$
 $\parallel \parallel \hat{I} \hat{d} \hat{T} \hat{I} \hat{S} \parallel \circ \parallel \hat{C} \hat{z} \hat{S} \hat{u} \parallel \hat{S} \hat{S} \parallel \hat{S} \hat{T} \hat{P}$

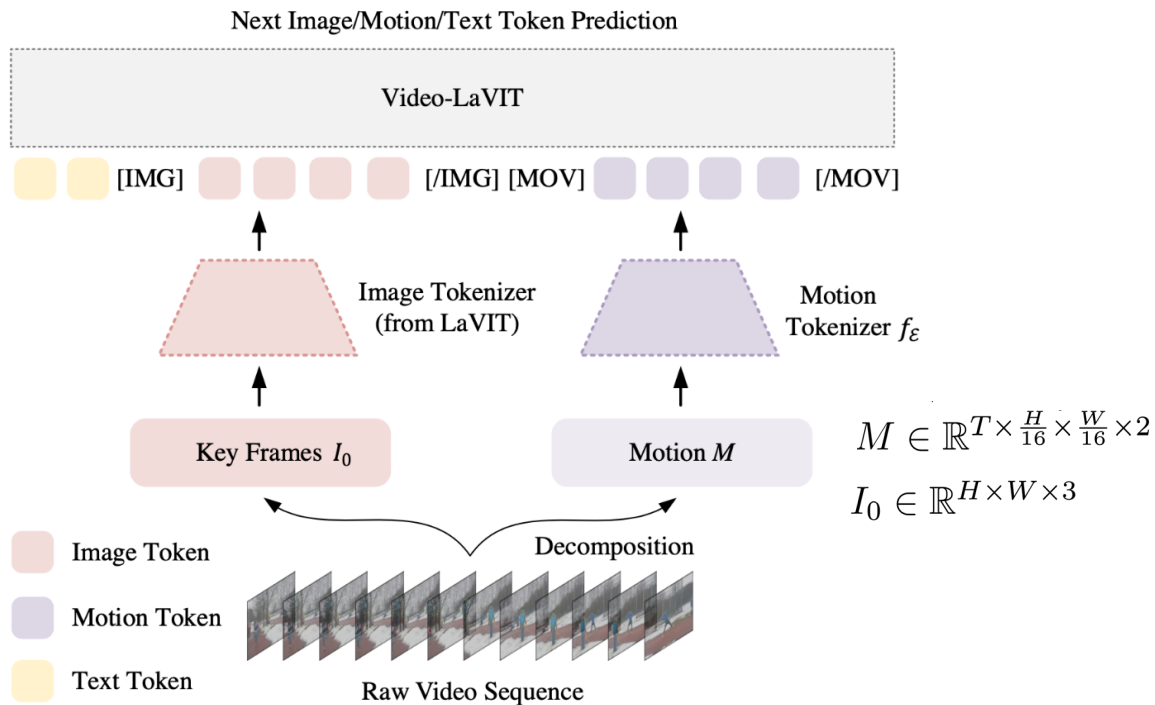


$$\vec{m}(p, q) = \arg \min_{i, j} \|I_t(p, q) - I_{t-1}(p - i, q - j)\|$$

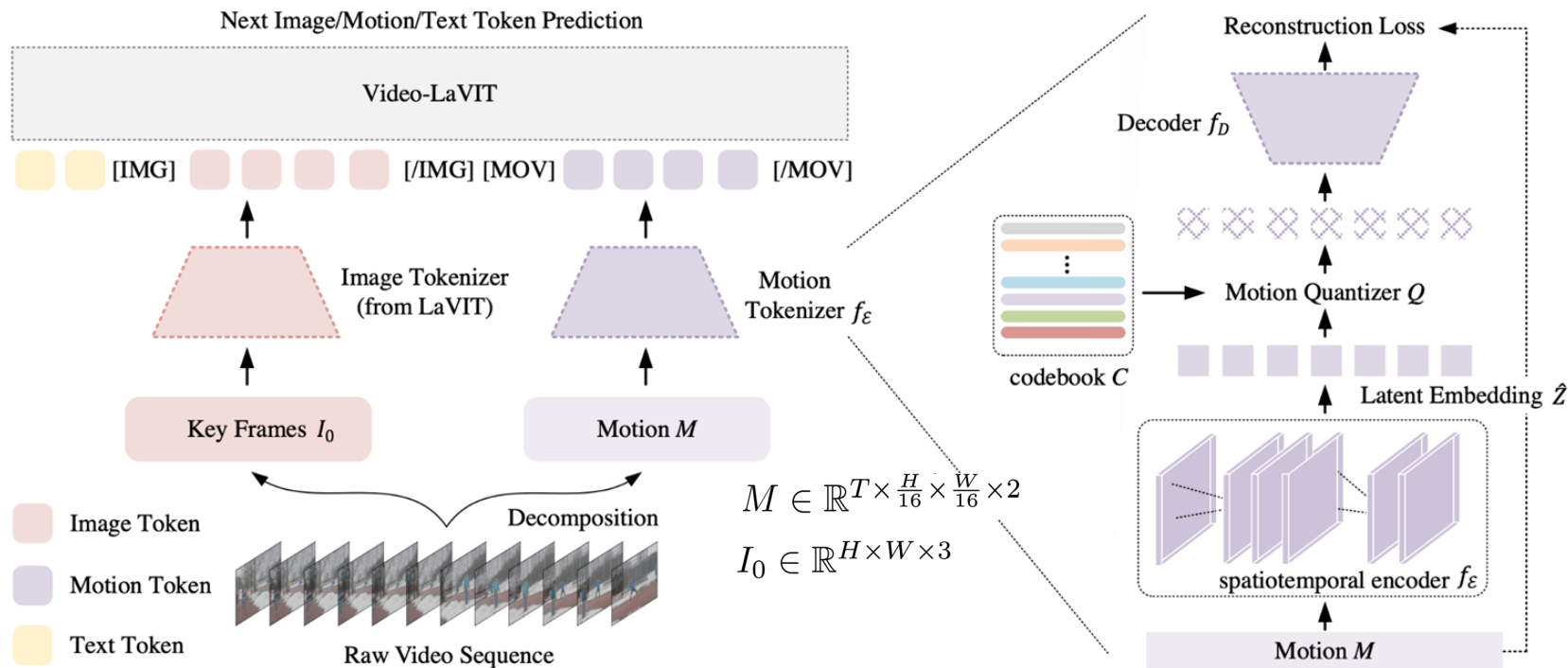
(i, j) K $\parallel \parallel \hat{G} \hat{e} \hat{T} \hat{I} \hat{S} \parallel \hat{Y} \hat{E} \hat{S} \hat{H} \hat{S} \hat{I} \hat{U} \hat{S} \hat{S} \hat{T} \hat{H} \uparrow \hat{S} \hat{H} \hat{S} \hat{T} \hat{S} \hat{C} \hat{E}$
 $\parallel \parallel \hat{E} \hat{E} \parallel \frac{1}{2} \hat{C} \hat{I} \parallel \parallel L \hat{T} \hat{E}$



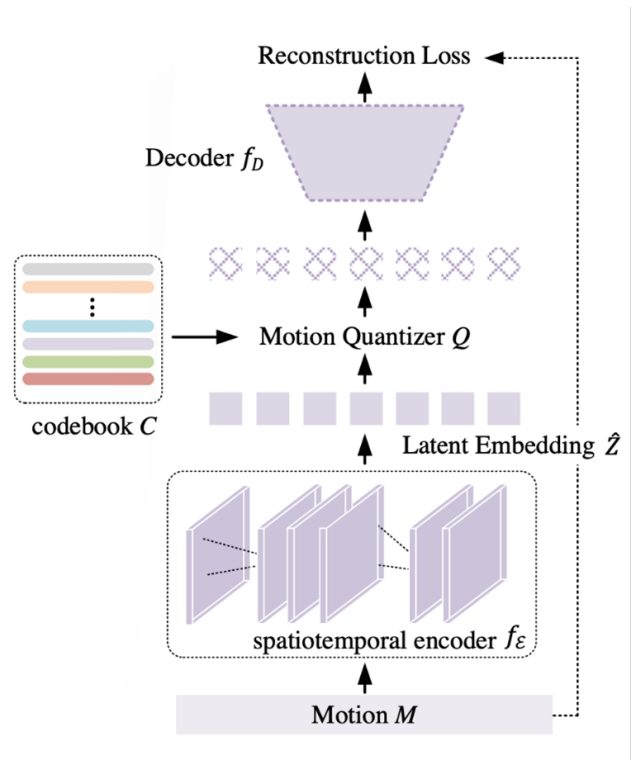
◦ $\hat{S} \parallel 3U_{1/2} \circ p \in \mathbb{H}^p; \hat{e} \in \mathbb{U}^E, \dots \text{EIEU} \hat{C}_{1/2} \mathbb{P}$



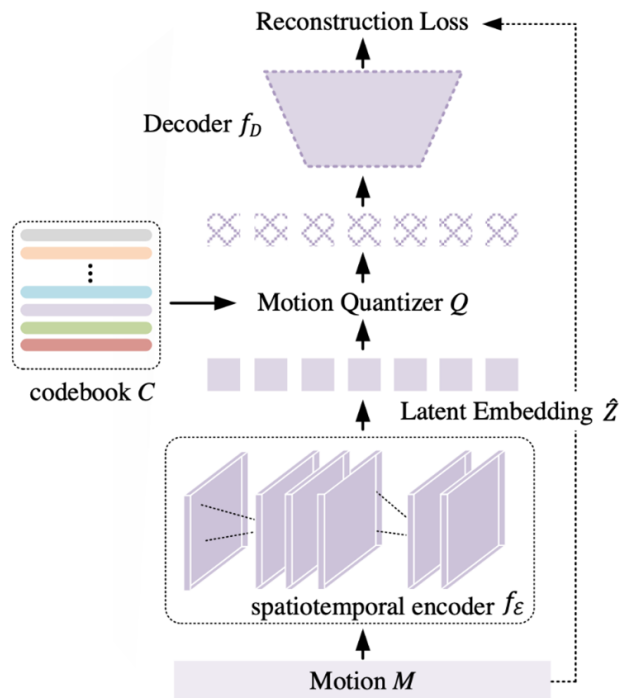
◦ $\hat{S} \parallel 3U_{1/2} \circ p \in \mathbb{H}^p; \hat{e} \in \mathbb{U}^E, \dots \text{EIEU} \hat{C}_{1/2} \mathbb{P}$



◦ $\hat{M} \hat{S} \parallel 3U^{1/2} \circ p \in \mathbb{H}^p; \hat{e} \in \mathbb{U}^E, \dots \hat{E} \hat{U} \hat{C} \hat{2} \hat{P}$



◦ $\hat{S} \parallel 3U_{1/2} \circ p \in \mathbb{H}^p; \hat{e} \in \mathbb{U}^e, \dots \text{EIEU } \hat{C}_{1/2} \mathbb{P}$



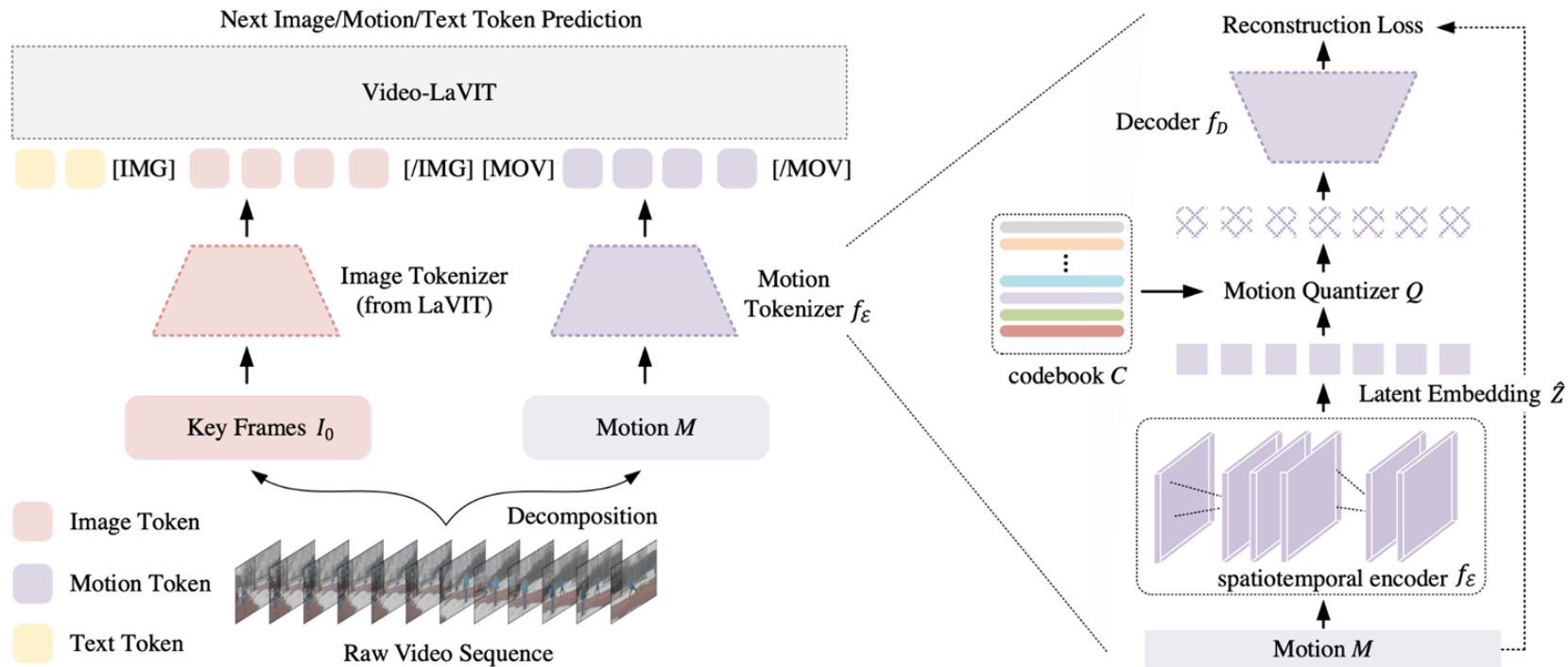
Each embedding vector is then tokenized by a vector quantizer

$$z_i = \arg \min_j \|l_2(\hat{z}_i) - l_2(c_j)\|_2$$

$$\hat{Z} \in \mathbb{R}^{N \times d}$$

$$M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$$

◦ $\hat{S} \parallel 3U_{1/2} \text{ p} \in \mathbb{H} \text{p}; \hat{e} \text{ Ü} \ddot{\text{E}}, \text{ " ' E I } \ddot{\text{U}} \hat{C}_{1/2} \text{ P}$



A

œ € ě || ■ ņ š ě

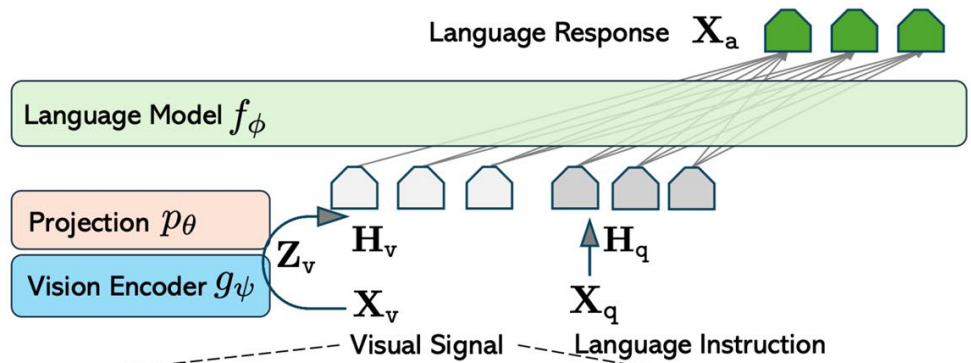
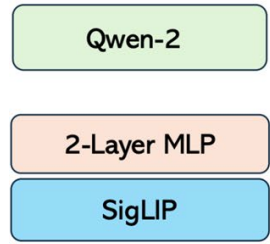
Some state -of-the-art open
source models on Video-MME
benchmark





Qwen2-VL: A Vision-Language Model for Video

- Qwen2-VL 3B 720P 16FPS 1080P 30FPS 4K 60FPS 8K 60FPS
- Qwen2-VL 72B 720P 16FPS 1080P 30FPS 4K 60FPS 8K 60FPS



Single Image



Multi-Image



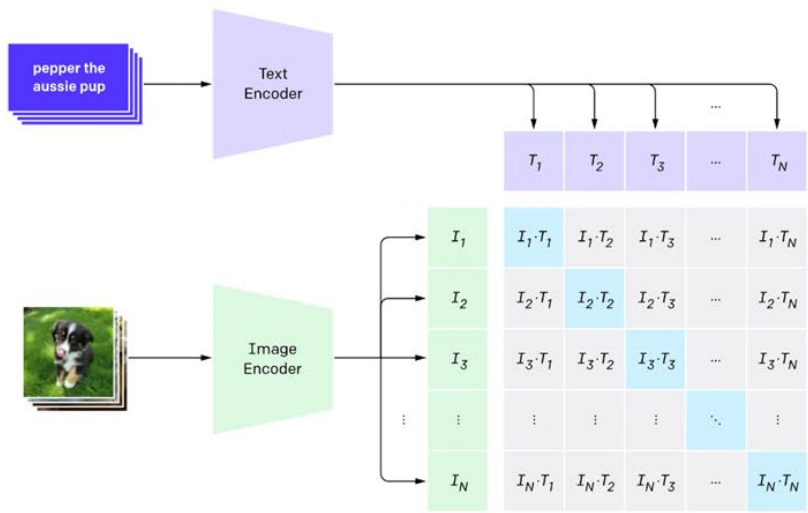
Video

— ĆŠ ¼|| Ğ½ĈÿË ĩp|| ½ⁿ ŠĬ=Ĭ ħ ŠĈËË

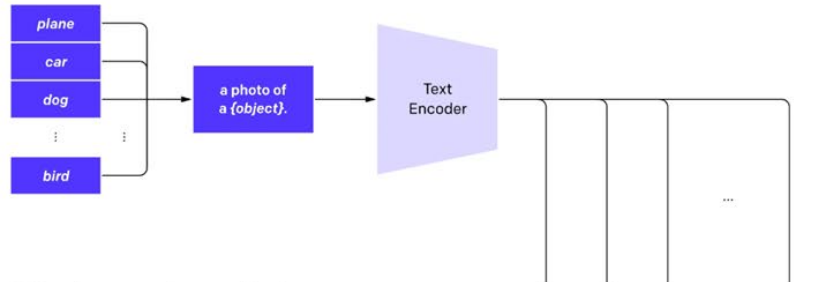
≈s h a l yš. □
 §l a y b k j h a s b l i a b
 ÷ ≈ ≈ ÷ _

CLIP: 400M (image, text) pairs, 500×V100 GPUs for pretraining

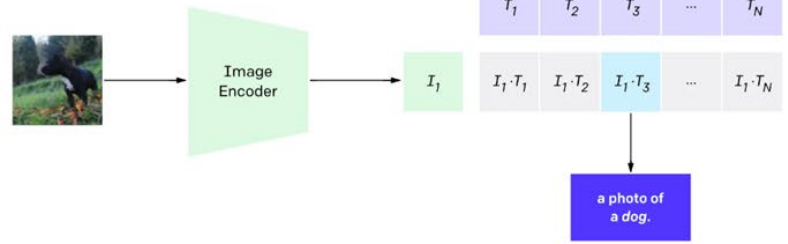
1. Contrastive pre-training



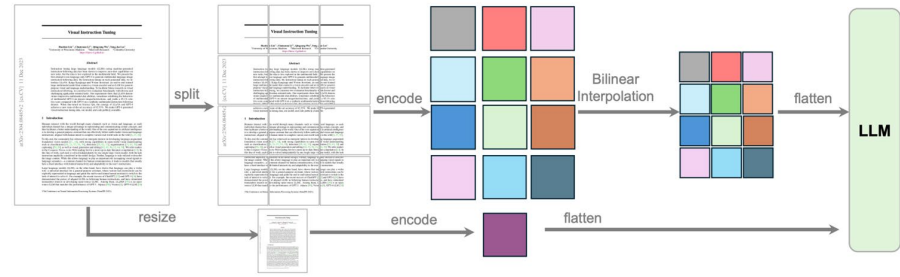
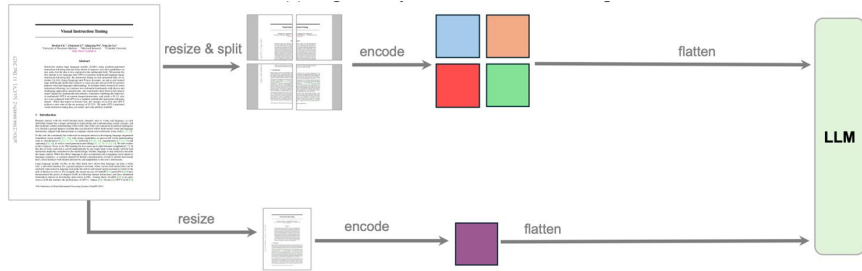
2. Create dataset classifier from label text



3. Use for zero-shot prediction



ÜÜ½ 3û ƒŠ° ΔE | ƒi k Δ ↓ E S B | ƒi Δ ƒ



Original AnyRes (“Max -9”)

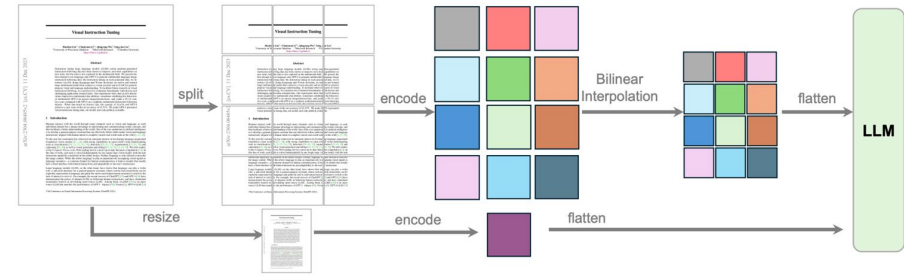
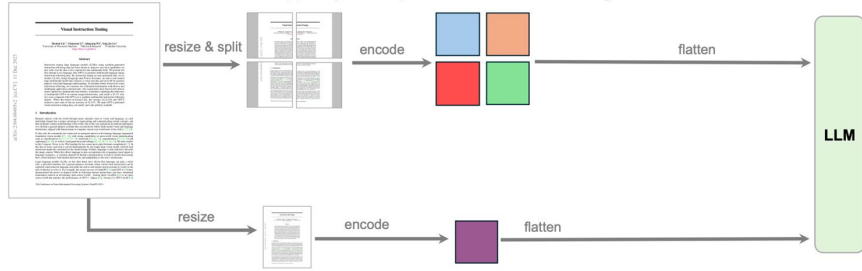
Split into 9 patches → each into 384² → patch 14 x 14

???

k Δ ↓ Š Ć E ƒ Ÿ ~ Š E H Δ ƒ S ½ Ć P ƒ Š Ć | ƒ i Δ ƒ



ÜÜ½ 3û ƒŠ° ΔÉ ƒİ k Δ †Ë Šİ ƒİΔ ƒ



û Ç ƒ½É ƒŸ~ŠÉİİÉ ½ŠİİÉ

Ç ƒİƒİ İÇÉ ½İL†ŠÉÉ → Š½L†İƒİ İAYEİÉ →
 • ½İL†İÉİİÉİÉ → Ö İİ ƒŠÉÉ →

İ ½İİÖ İİ ƒŠÉÉ

k Δ †ŠÇÉ ƒŸ~ŠÉİİ ΔŠ½ÇƒİŠÇ ƒİΔ ƒ

pİİ†ŠİÇÜİİ ƒŠƒİ İİŠŸŠŠİİİ†Šİ†ŠÉİ ƒİÉ
 ½İÉ ½İL†ŠÉÉ/Šİİ ÜƒÉ½ • ŠİİΔİİ ΔŠ½ÇÉ
 ƒİŠÇ ƒİΔ ƒİ İŠŸƒİŠİİ İŠİİ ƒŠÉÉ ŠÇİÇ •



ÜÜ½ 3û ₪° Æ| ₪k Æ ↓ ÆŚ| ₪| ₪

729 is the #tokens for SigLIP to encode a visual input of resolution 384 × 384

 <p>Single-Image</p>	 <p>... N Crops</p>	<p>$(1 + 9) * 729 = 7290$ Tokens</p>
<p>729 + N * 729 Tokens</p>		
 <p>Multi-Image</p>	 <p>... N Images</p>	<p>$12 * 729 = 8748$ Tokens</p>
<p>N * 729 Tokens</p>		
 <p>Video</p>	 <p>... N Frames</p>	<p>$32 * 196 = 6272$ Tokens</p>
<p>N * 196 Tokens</p>		
<p>Example on Token Strategy</p>		<p>Max Tokens</p>



ÜÜ½° 3û ₪° ∆E∣ =İİ₪₪!ŚĈŚ₪LŚ

S5: Video-to-Video Difference



User Here are two videos, please provide the analysis on their differences.

LLaVA-OV The first video shows a sequence of images where a man in a blue plaid shirt approaches a woman in a red dress, engages in a conversation, and eventually gives her a kiss on the cheek. The woman in the red dress is looking at the man with a slight smile, and another woman in a light blue top is observing the scene. The setting appears to be a street with buildings in the background. The second video shows the same man in the blue plaid shirt walking away from the woman in the red dress, who is now looking in the opposite direction. The woman in the light blue top is no longer visible in this sequence. The man is walking down the street, passing by the buildings and other pedestrians. The overall mood of the video changes from the intimate and personal interaction in the first video to a more solitary and neutral scene in the second video.

S3: Set-of-mark Prompting



User Here's an image with objects labeled in numbers. Please describe the objects that are marked as 4, 5 and 7.

LLaVA-OV The image shows a room with various items marked with numbers. Here's the description for the items marked as 4, 5, and 7:
The item marked as 4 is a framed picture hanging on the wall. It features a black and white illustration of a dandelion with seeds blown away, creating a sense of movement. The item marked as 5 is a white bookshelf with multiple shelves. It appears to be a storage unit that is part of the room's furniture. It is located between the two desks. The item marked as 7 is a black chair with a modern design. It has a metal frame and a woven seat, suggesting a contemporary style. This chair is positioned in front of one of the desks.

ÜÜ½° 3û 𐌹° 𐌲𐌸 𐌹→ ÜÜ½° 3° 𐌲 𐌹

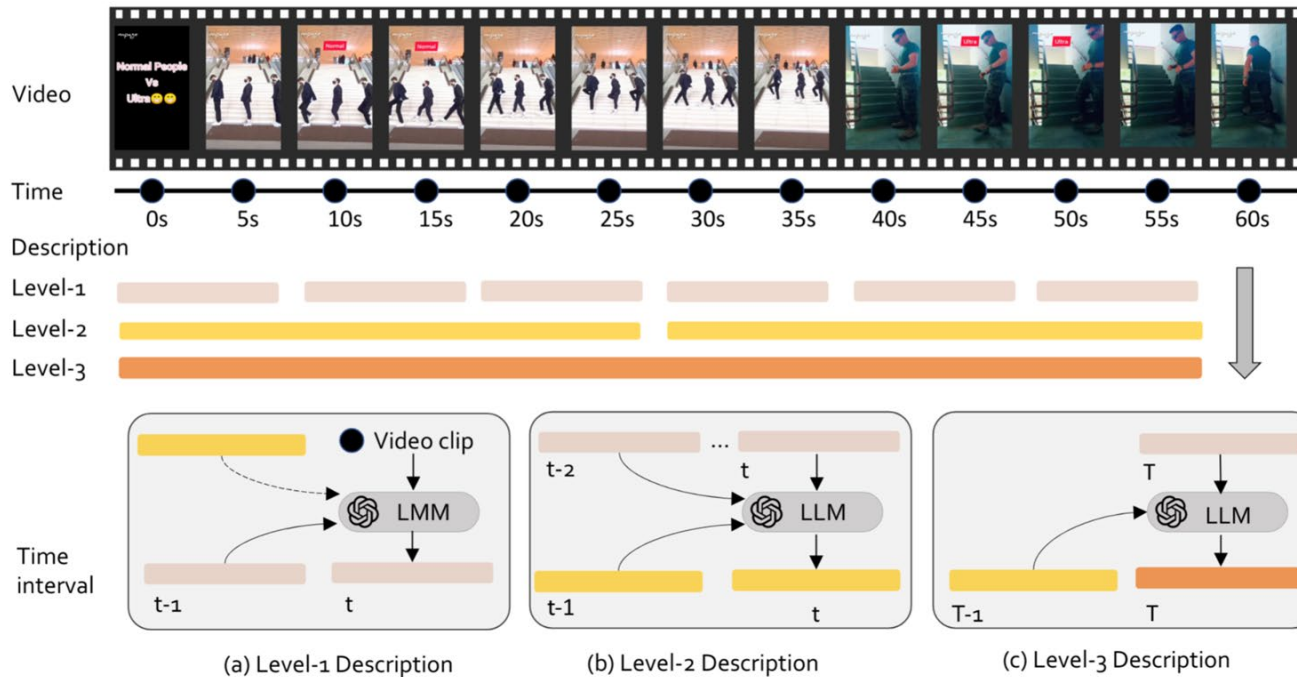
k | Ü 𐌸 | 𐌸 ½↕S 𐌸 𐌹↕S 𐌸 𐌹 ½↕ ½↕



Source	Filtering Logic
HD-VILA-100M	①②③④
InternVid	①②③④
ActivityNet	③⑤
VIDAL	①②③④
Charades	③⑤
Ego4D	③⑤
VidOR	③⑤
YouCook2	③⑤⑥
Kinetics-700	③⑤⑥
Sth-sthv2	③⑤⑥

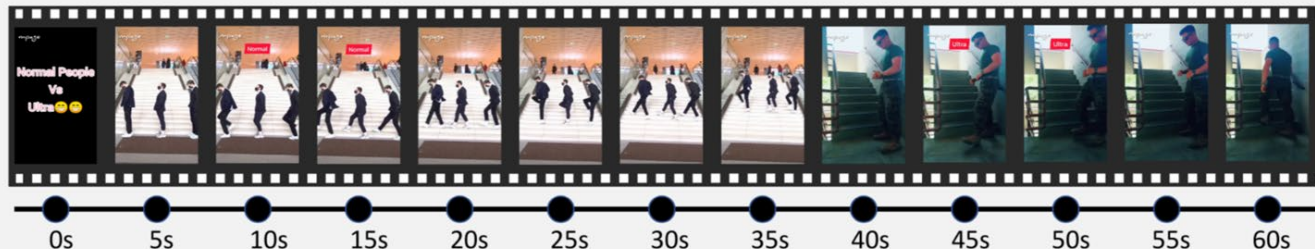
- 10 datasets from which all are inherited
- Filtering steps (long video, diverse, high-quality)
- In total is 178k videos

ÜÜ½° 3û =Š° ΔE|| =E → ÜÜ½° 3° ãŠ||



GPT has finite length of input context, so authors came up with a **recursive LLM scheme**

ÜÜ½° 3û Ɔ° ΔEΔ | ƆE → ÜÜ½° 3° ð Š |



Annotation type 1: detailed description

The video begins with a black screen displaying the text 'Normal People Vs Ultra' in pink and white letters, accompanied by two smiling face emojis. The scene transitions to a modern building with a staircase. Three individuals, dressed in black suits and white sneakers, stand in a line on the stairs. The text 'Normal' appears in a red box at the top left corner. The individuals start walking up the stairs in a synchronized manner, maintaining their formation. The background shows a few people walking and an escalator on the right side of the stairs. The individuals continue to walk up the stairs in a coordinated manner. The scene then transitions to... <omited>

Annotation type 2: open-ended question

Question: How many steps does "normal people" climb?
 Answer: "Normal people" climb 7 steps in the video.

Annotation type 3: multi-choice question

Question: How many steps does "normal people" climb? A. 5 B. 6 C. 7 D.8
 Answer: C.7

ÜÜ½° 3û 𐌆° 𐌆𐌆 𐌆𐌆 → ÜÜ½° 3° 𐌆𐌆

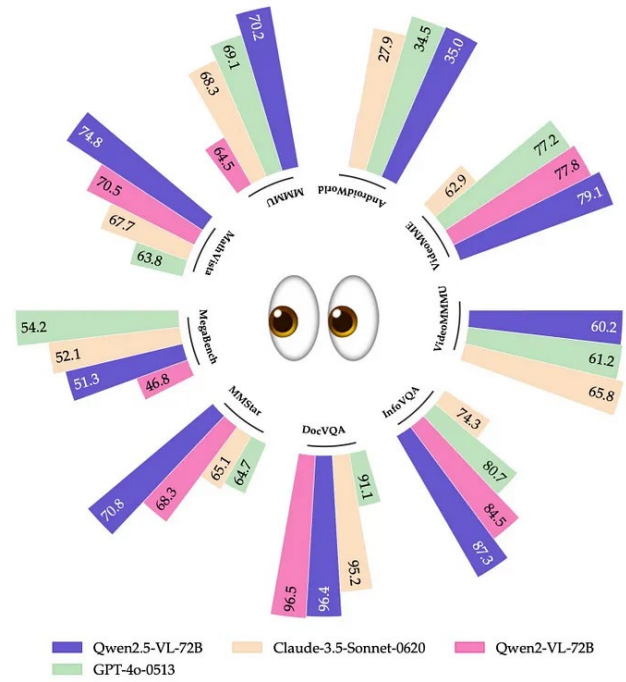
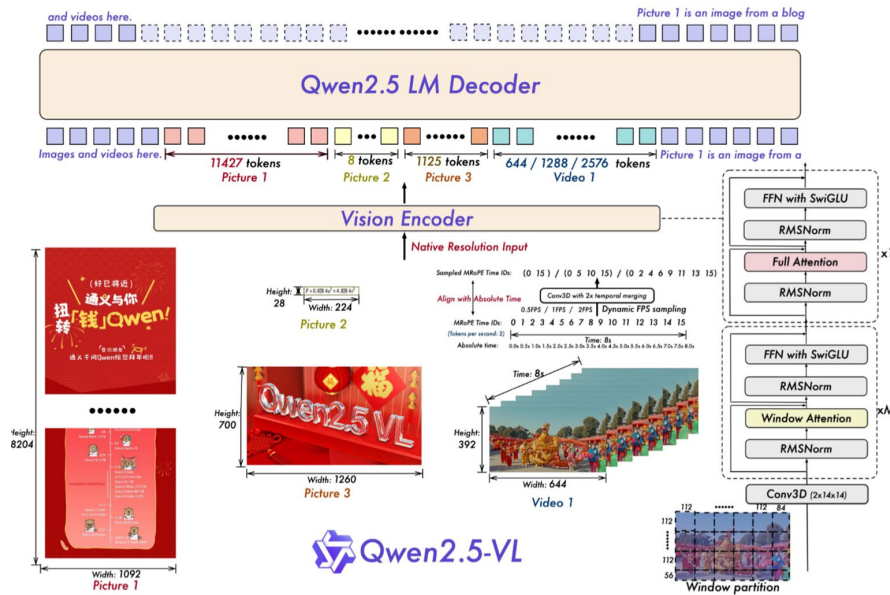
pĩĤ½/𐌆ĩ ŠĤĖŠ°𐌆ĩ↓½ĩĖ
 I 𐌆 𐌆 𐌆 ½ŠñĤ Ĥ ĤÜ½° 3
 û 𐌆° 𐌆𐌆 𐌆𐌆Ü½° 3° 𐌆𐌆
 Ũ 𐌆ĩĤĩ↓ŠĤ Ĥ½ĩ½ĩŠ 𐌆 ĤĤ

Model	Caption		Open-Ended Q&A		Multi-Choice Q&A						
	VideoDC	Dream-IK	ActNet-QA	VideoChatGPT	EgoSchema	MLVU	MVBench	NExT-QA	PerceptionTest	Long VideoBench	VideoMME
	test	test	test	test	test	m-avg	test	mc	val	val	wo/w-sub
<i>Proprietary models</i>											
GPT-4V (OpenAI, 2023)	4.00	34.4	57.0	4.06	-	49.2	43.5	-	-	61.3	59.9/63.3
GPT-4o (OpenAI, 2024)	-	39.2	-	-	-	64.6	-	-	-	66.7	71.9/77.2
Gemini-1.5-Flash (Team et al., 2023)	-	34.8	55.3	-	65.7	-	-	-	-	61.6	70.3/75.0
Gemini-1.5-Pro (Team et al., 2023)	-	36.2	57.5	-	72.2	-	-	-	-	64.0	75.0/81.3
<i>Open-source models</i>											
VILA-40B (Lin et al., 2024)	3.37	33.2	58.0	3.36	58.0	-	-	67.9	54.0	-	60.1/61.1
PLLaVA-34B (Xu et al., 2024a)	-	28.2	60.9	3.48	-	-	58.1	-	-	53.2	-
LongVA-7B (Zhang et al., 2024c)	3.14	-	50.0	3.20	-	56.3	-	68.3	-	-	52.6/54.3
IXC-2.5-7B (Zhang et al., 2024b)	-	-	52.8	3.46	-	37.3	69.1	71.0	34.4	-	55.8/58.8
LLaVA-OV-7B (Li et al., 2024c)	3.75	31.7	56.6	3.51	60.1	64.7	56.7	79.4*	57.1	56.5	58.2/61.5
VideoLLaMA2-72B (Cheng et al., 2024)	-	27.1	55.2	3.16	63.9	61.2	62.0	-	-	-	61.4/63.1
LLaVA-OV-72B (Li et al., 2024c)	3.60	33.2	62.3	3.62	62.0	68.0	59.4	80.2*	66.9	61.3	66.2/69.5
LLaVA-Video-7B	3.66	32.5	56.5*	3.52	57.3	70.8	58.6	83.2*	67.9*	58.2	63.3/69.7
LLaVA-Video-72B	3.73	34.0	63.4*	3.62	65.6	74.4	64.1	85.4*	74.3*	61.9	70.5/76.9



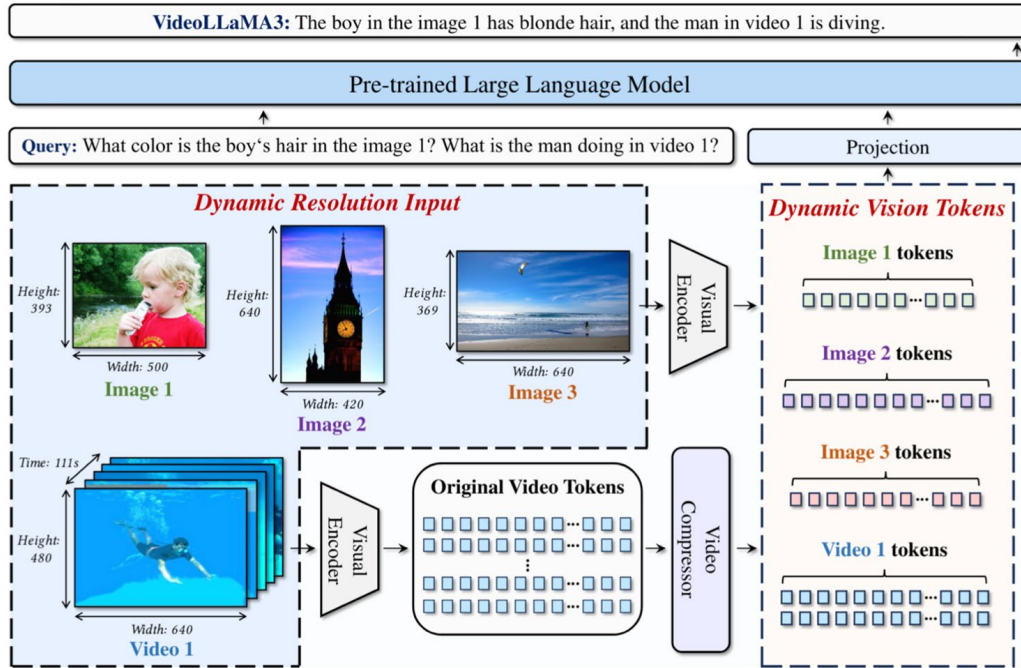
ƒ ŮŚƚ SB° ÜΠ ℔ ½ ½Ë, ƚHP

° Æ ƚ ƚ ñ Ś Ć Ć ½ ½ ñ Ć || Ć Ć Ć Ć










◦ ñ Š | ÜÜ½æ ÄË Ć↕ ãŚLĪĴ ĆŠĤ , “HP



Dynamic Resolution Input

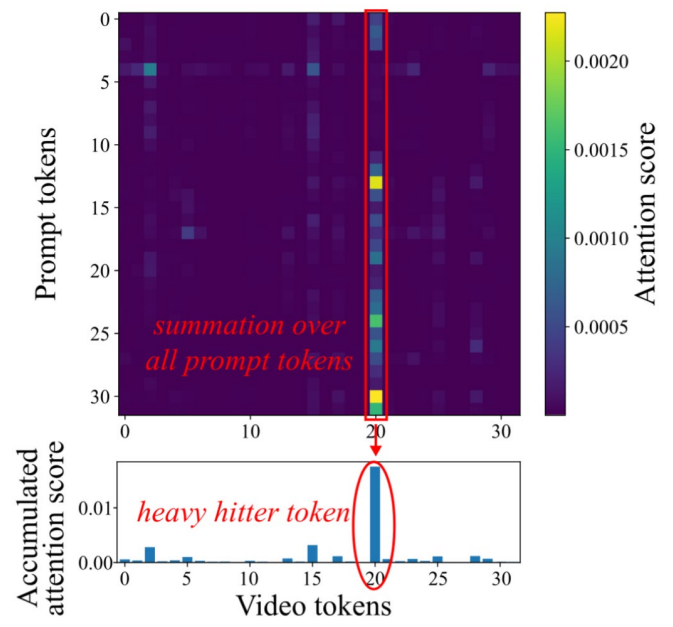
SigLIP has been fine-tuned to accept images of different sizes, not only 336 x 336

° ħ Ś | ÜÜ½æ ÄËLū½Ɔ½Ĩ | =

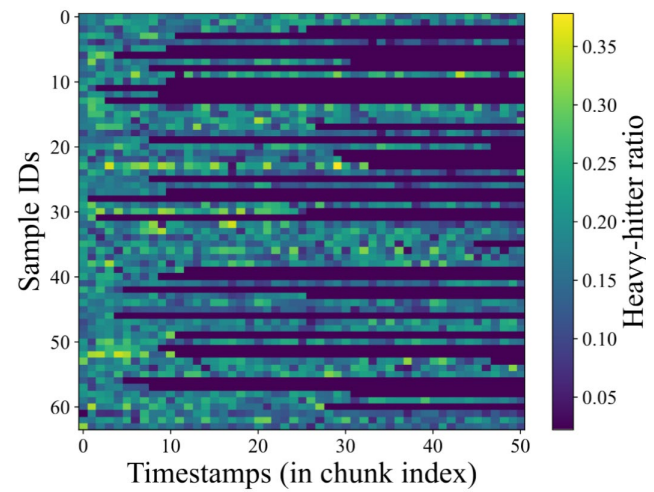
	Qwen2-VL 7B 	InternVL2.5 8B 	LLaVA-Video 7B 	NVILA 8B 	Apollo 7B 	VideoLLaMA 2.1-7B	VideoLLaMA 3-7B
<i>General Video Understanding</i>							
VideoMME <i>w/o sub</i>	63.3	<u>64.2</u>	63.3	<u>64.2</u>	61.3	54.9	66.2
VideoMME <i>w/ sub</i>	69.0	66.9	69.7	<u>70.0</u>	63.3	56.4	70.3
MMVU _{val}	42.1 [†]	41.1 [†]	42.4*	<u>43.7*</u>	-	39.5 [†]	44.1
MVBench	67.0	72.0	58.6	68.1	-	57.3	<u>69.7</u>
EgoSchema _{test}	66.7	<u>66.2*</u>	57.3	54.3*	-	53.1	63.3
PerceptionTest _{test}	62.3	68.9*	<u>67.9*</u>	65.4*	-	54.9	72.8
ActivityNet-QA	57.4*	58.9*	56.5	<u>60.9</u>	-	53.0	61.3
<i>Long Video Understanding</i>							
MLVU _{dev}	69.8*	69.0*	70.8*	70.6*	<u>70.9</u>	57.4	73.0
LongVideoBench _{val}	55.6 [†]	60.0	58.2	57.7	58.5	-	<u>59.8</u>
LVBench	<u>44.7*</u>	43.2*	41.5*	44.0*	-	36.2	45.3
<i>Temporal Reasoning</i>							
TempCompass	67.9 [†]	<u>68.3*</u>	65.4	69.7*	64.9	56.8	68.1
NextQA	81.2*	85.0*	83.2	82.2	-	75.6	<u>84.5</u>
Charades-STA	-	-	-	-	-	-	60.7

ñ 1/2 ≈ S € 1/2 π S

k S 1/2 n Y 3 t n S C C 1 d I I E S 1/2 E S C S n J n 1/2 L Y

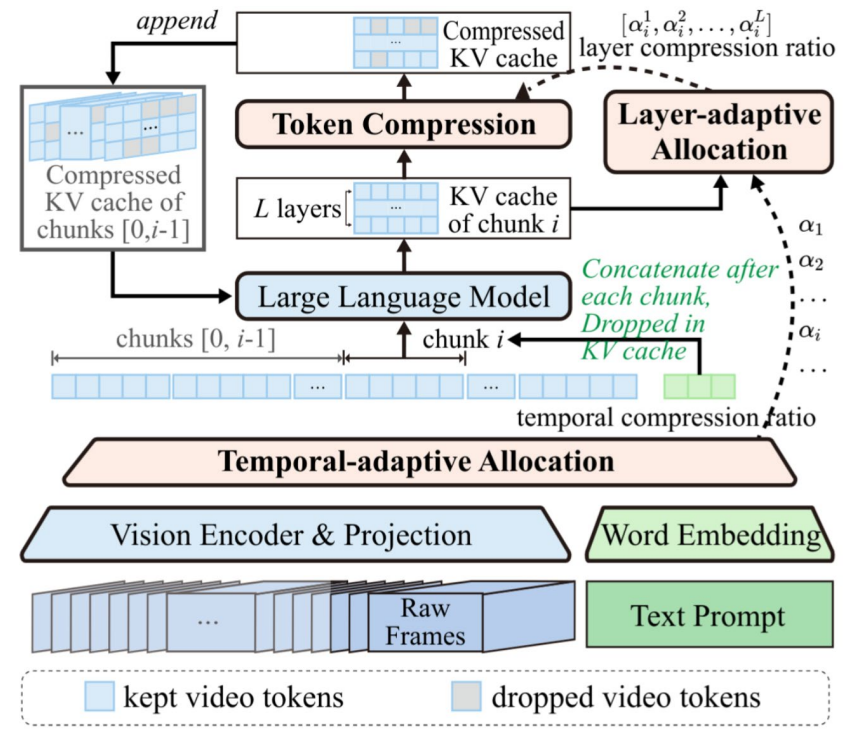


Redundancy among video timestamps



$$\tilde{n}^{1/2} \approx \hat{S} \epsilon^{1/2} \Pi \hat{S}$$

- Measure temporal redundancy in chunks of video
- Chunks with **low change** (high similarity) are considered redundant and receive **heavier compression**
- **No retraining** : All decisions use computed similarities and attention scores



ñ 1/2 ≈ S€ 1/2 π S i E L ū 1/2 ♯ 1/2 ã □ =

Model	LLM Size	VideoMME		MLVU	LongVideoBench		LVBench
		Long	Overall	dev	val	val	
GLM-4V-Plus	-	-	70.8	-	-	-	58.7
GPT-4o	-	65.3	71.9	64.6	66.7	66.7	27.0
Gemini-1.5-Pro	-	67.4	75.0	-	64.0	64.0	33.1
VITA-1.5	7B	47.1	56.1	-	-	-	-
mPLUG-Owl3	7B	50.1	59.3	63.7	52.1	52.1	-
NVILA	8B	54.8	64.2	70.1	57.7	57.7	-
ByteVideoLLM	14B	56.4	64.6	70.1	-	-	-
TPO	7B	55.4	65.6	71.1	60.1	60.1	-
VideoLLaMA3	7B	-	66.2	73.0	59.8	59.8	45.3
LLaVA-Video	7B	52.4	63.3	67.0	58.2	58.2	43.1
LLaVA-Video+AdaRETAKE	7B	53.9	64.0	70.6	59.6	59.6	49.6
Qwen2-VL	7B	53.8	63.3	66.9	55.6	55.6	42.4
QWen2-VL+AdaRETAKE	7B	56.4	64.2	72.0	57.2	57.2	48.9
Qwen2.5-VL	7B	55.6	65.4	70.2	59.5	59.5	45.3
QWen2.5-VL+AdaRETAKE	7B	58.3	67.7	75.0	62.6	62.6	51.2
LLaVA-One Vision	72B	60.0	66.3	68.0	61.3	61.3	-
Oryx-1.5	32B	59.3	67.3	72.3	62.0	62.0	30.4
Aria	8x3.5B	58.8	67.6	70.6	65.3	65.3	-
LLaVA-Video	72B	61.5	70.6	74.4	61.9	61.9	-
Qwen2-VL	72B	62.2	71.2	-	60.4	60.4	41.3
InternVL2.5	72B	62.6	72.1	75.7	63.6	63.6	43.6
Qwen2.5-VL	72B	63.9	72.6	74.6	65.9	65.9	47.3
Qwen2.5-VL+AdaRETAKE	72B	65.0	73.5	78.1	67.0	67.0	53.3

- , 1/2 E E f S n C A S n E S u S C 2 E E I 1/2 S 3 ! 3 I 1/2 S 3 1/2 C E U U e E I E I L 1/2 E 1/2 E E , S T 3 U E 1/2 n E U 1/2 3 ° m S
- L 1/2 S u S n E | . C u S | S T E I ! E ~ S A E 1/2 n E S Y | T E I 1/2 n 1/2 G n E I 1/2 E T E
- d 1/2 n E I ! E H Q E 1/2 n E S | | T E 1/2 S E U : S T 1/2 E 1/2 E S I I U 1/2 L 1/2 E L | | . C E S n E T S C u m S E

$\hat{Y}3\tilde{I} \blacksquare 3 \hat{Y}\ddot{E} \blacksquare \blacksquare \check{n} \frac{1}{2} \mathbb{L} \hat{Y}$

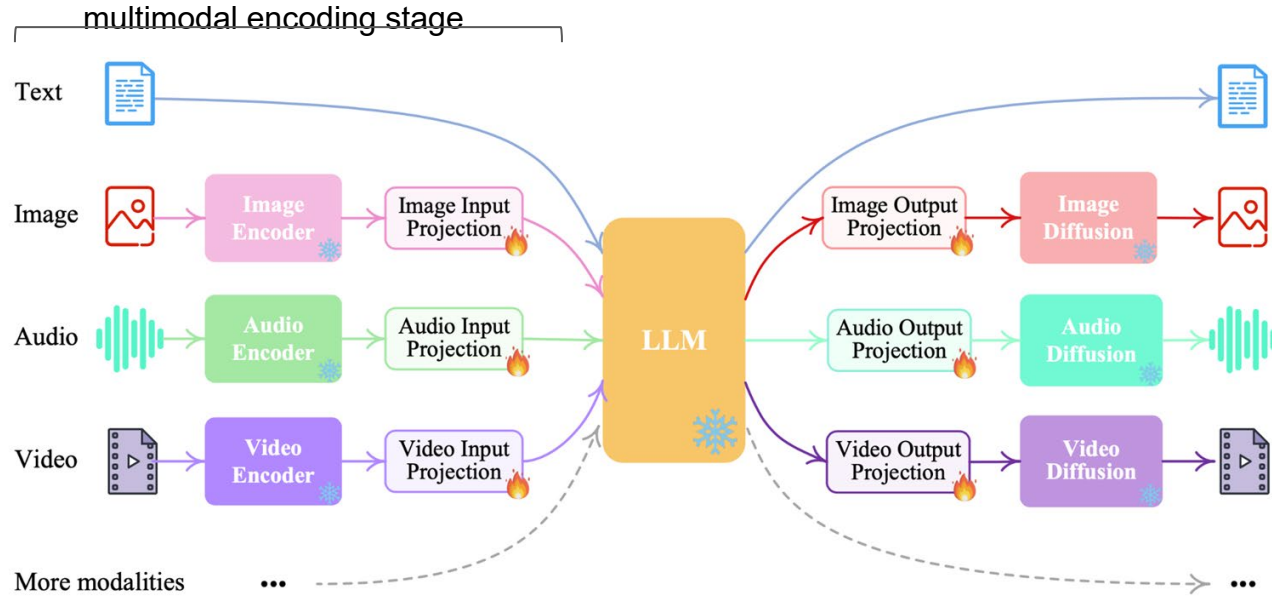
$L \frac{1}{2} \hat{C} \mathbb{L} \hat{Y} \hat{E} \hat{J} \hat{E} \blacksquare \hat{I} \blacksquare \circ \acute{S} \hat{E} \tilde{H} \updownarrow \acute{S} \hat{E}$

$\check{n} \blacksquare \blacksquare \hat{C} \hat{E} \blacksquare \hat{C} \hat{E} \frac{1}{2} \hat{Y} 3 \tilde{I} \blacksquare 3 \frac{1}{2} \hat{Y} \hat{E}$

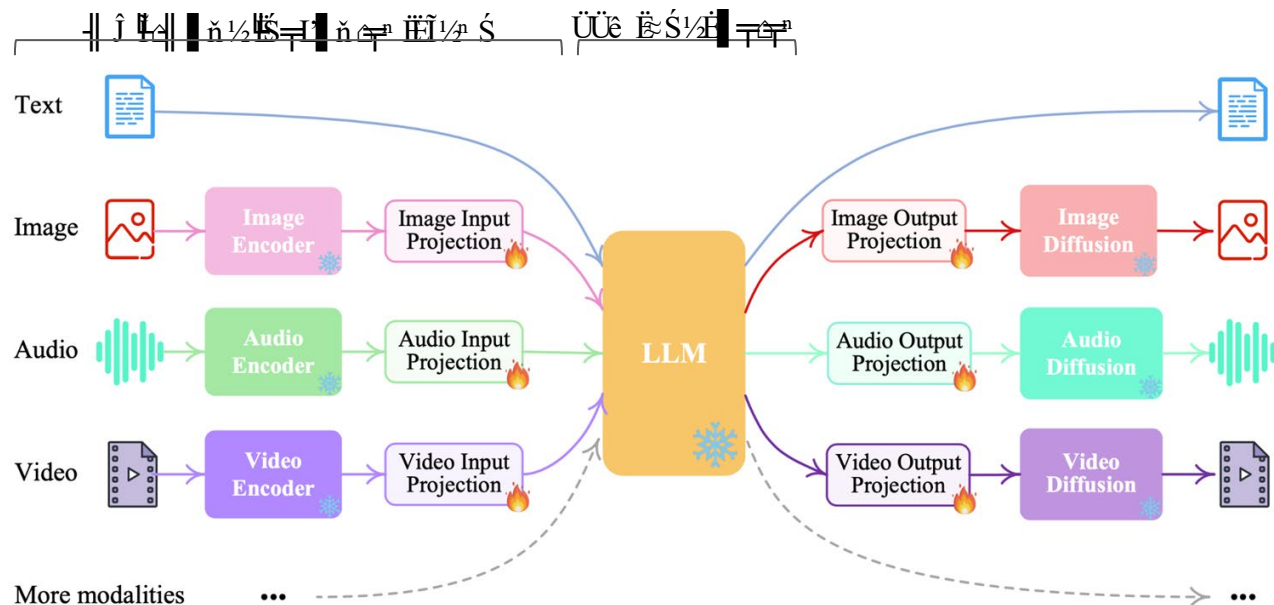
$\blacksquare \blacksquare \check{n} \frac{1}{2} \mathbb{L} \hat{Y}$



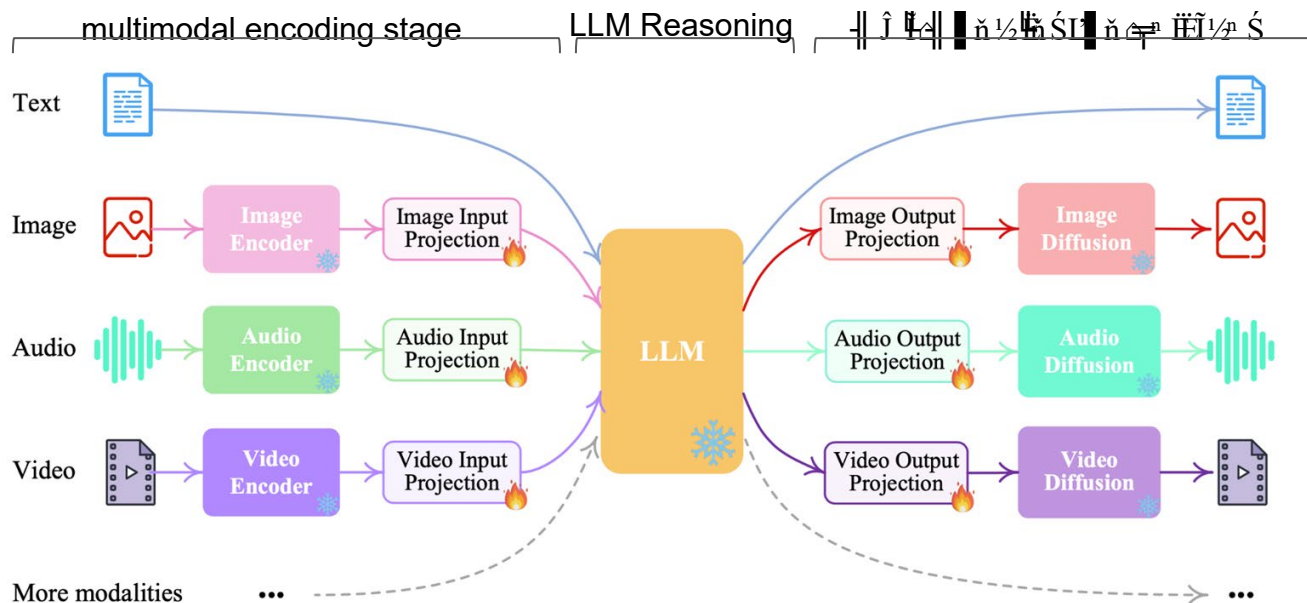
NEXT-GPT: Methodology



NEXT-GPT: Methodology



LLM Reasoning with Multimodal Encodings



í Lw€3d →€İ ê Šİ↕ █ ñ █ █ n ŸË

≈s h u yš. □
 | ʒ s ʒ ÷ ʒ k □ ʒ n



û → ŸË → ĴİË İ Ĵİ• ĴİË █ şLİ █ → ½ ŸŸĈËË/ŸŸĈŸa Ĵ ŸŸñ Ĥ █ Ĥ ŠĤ • ñ ½ ĴŸñ İŮ ã ↕ Ĥ ½ Ÿİ ↕ ŠİŸĤİË
 Š → █ ñ ŸĈËË/ñ Ĥ şLİ █ ñ ŸĈËË ŸŸ → Ĥ Ĥ Ĥ Ĥ Ĥ ½ Ĉ ½ █ ŠİŸĈĤİ | P → ŠŠñ Šñ Ĥ █ Ĥ ŠĤ • ñ ½ ĴŸñ

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	—	—	—	—	—	—
Image					Vicuna	7B ❄️	Transformer	31M 🔥	SD	1.3B ❄️
Audio	ImageBind	1.2B ❄️	Grouping	28M 🔥	(LoRA)	33M 🔥	Transformer	31M 🔥	AudioLDM	975M ❄️
Video							Transformer	32M 🔥	Zeroscope	1.8B ❄️

ImageBind: One Embedding Space To Bind Them All. CVPR 2023. FAIR, Meta AI. [\[link\]](#)

í Lw€3d -€İ p|| 1/2 Š: ĞñĚ

Only image -paired data is sufficient to bind the modalities together

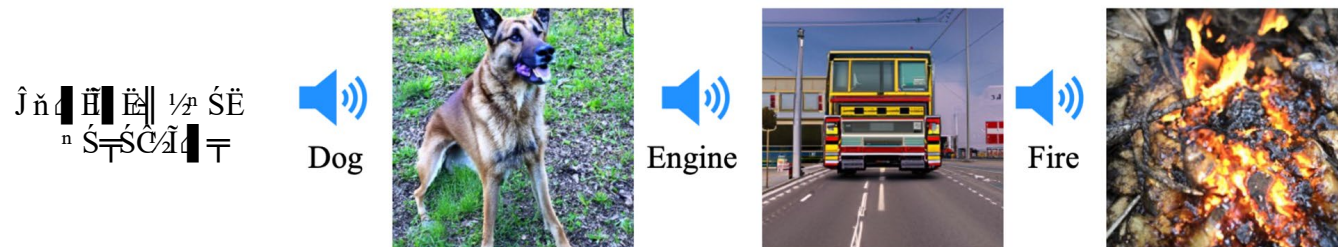
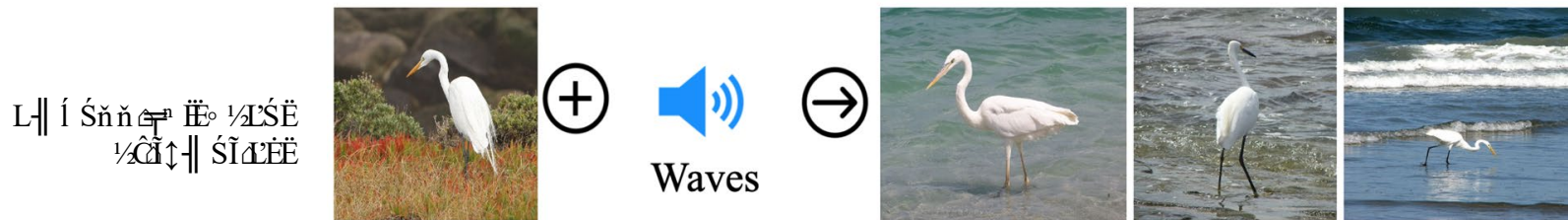


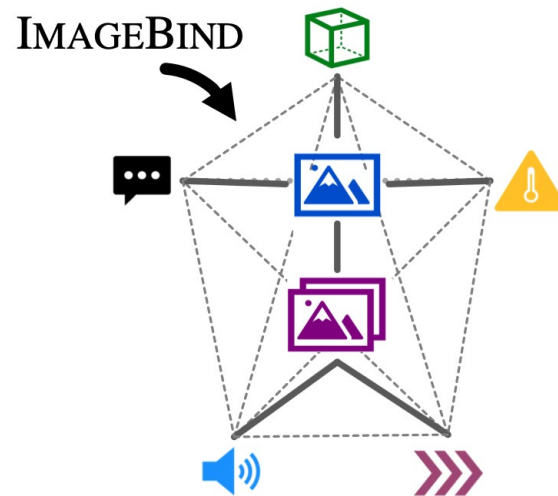
Image-Text Retrieval

Given a set of images \mathcal{I} and a set of queries \mathcal{M} , we want to find the most relevant image I_i for each query M_j .

$$\mathbf{q}_i = f(\mathbf{I}_i) \quad \mathbf{k}_j = g(\mathbf{M}_j)$$

$$L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^T \mathbf{k}_j / \tau)}{\exp(\mathbf{q}_i^T \mathbf{k}_j / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^T \mathbf{k}_j / \tau)}$$

$$(\mathcal{I}, \mathcal{M}_1) \text{ and } (\mathcal{I}, \mathcal{M}_2) \longrightarrow (\mathcal{M}_1, \mathcal{M}_2)$$



í Lw€3d -€i p|| 1/2 S: Ğñ È

≈s hλγ ŷš. □
| ⌘ š Ɔ ÷ Ɔk □ Ɔλn



L|| í Šññ Ğ Ę 1/2 Š È Ɔ λ γ ŷ š. □
| ⌘ š Ɔ ÷ Ɔk □ Ɔλn

Audio

Images & Videos

Depth

Text



Crackle of a Fire



“A fire crackles while a pan of food is frying on the fire.”

“Fire is crackling then wind starts blowing.”

“Firewood crackles then music...”



Baby Cooing



“A baby is crying while a toddler is laughing.”

“A baby is laughing while an adult is laughing.”

“A baby laughs and something...”

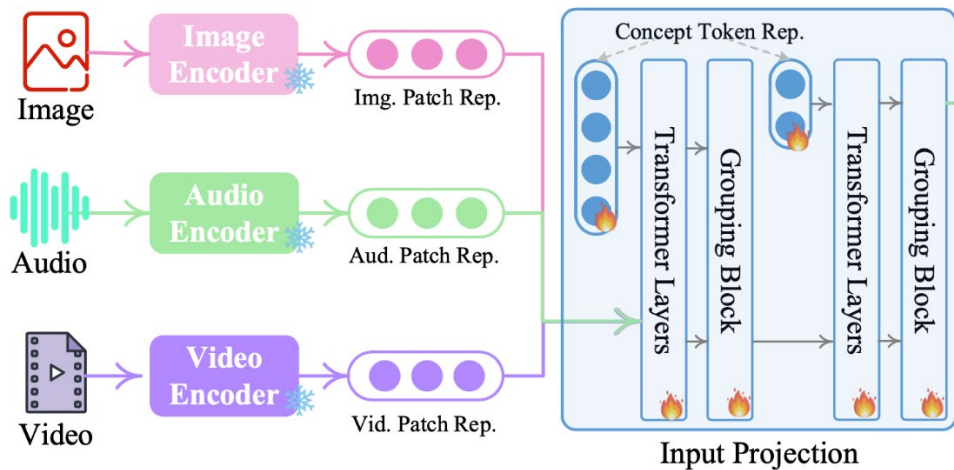
BindDiffusion: One Diffusion To Bind Them All. [\[link\]](#)

InternGPT: Solving Vision -Centric Tasks by Interacting with ChatGPT Beyond Language. [\[link\]](#)

í Lw€3d -€İ €Ĉ/234n Ę

Encoding Stage: Multimodal Alignment

B S E Δ T E / H Y ° S I ! E S 1 / G 1 / S E U / T L S ° İ H / T S T E M Ę Δ C / C ↑ L 1 / 2 W E / 2 n Ć n 1 / 2 İ S H ↓ S E Ć n 3 S u S E S 1 / J Ć S E Ę



can obtain $X^* = \{x_i^*\}_{i=1}^{N^*}$ multimodal tokens

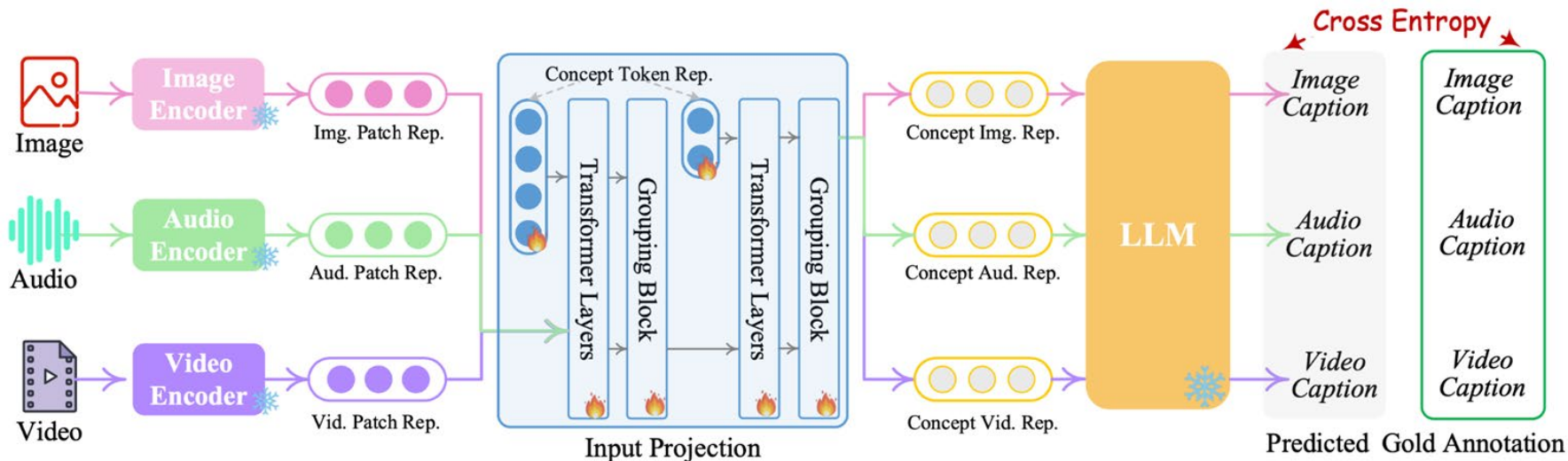
$$\hat{C}^l = \{c_j\}_j$$

$$\hat{C}^l, \hat{X}^l = \text{Transformer}([C^l; X^l])$$

Encoding Stage: Multimodal Alignment

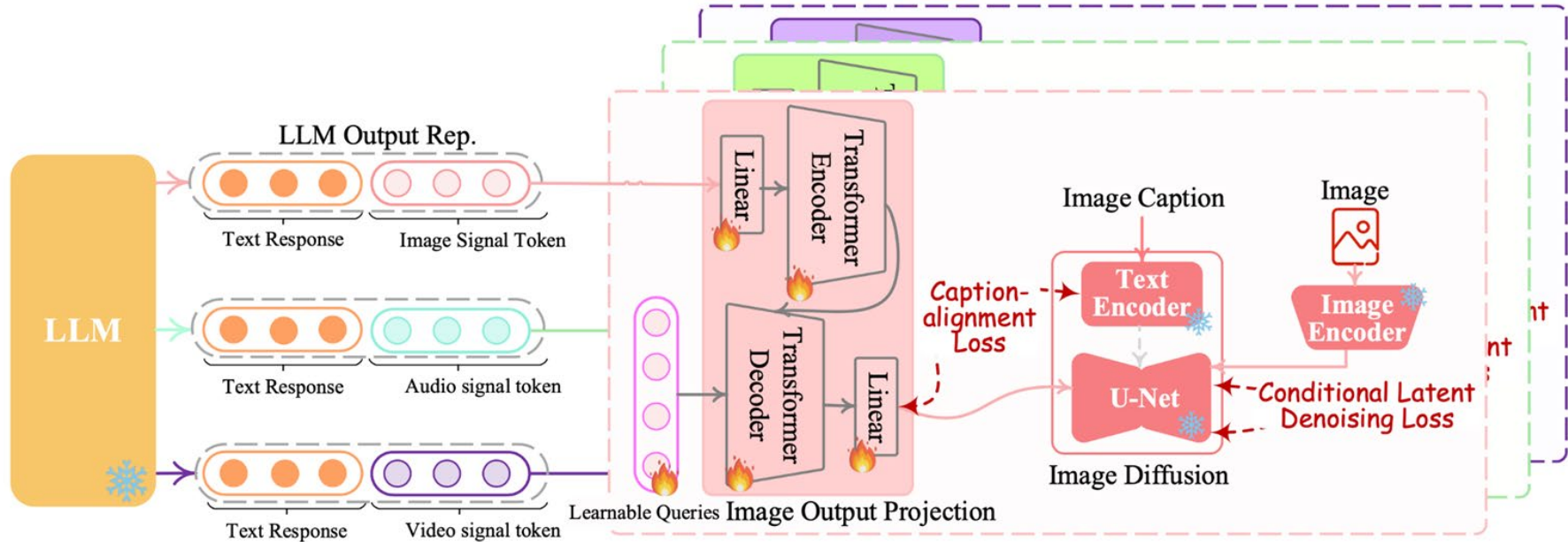
Encoding Stage: Multimodal Alignment

Diagram illustrating the encoding stage for multimodal alignment, showing the flow from input modalities (Image, Audio, Video) through encoders and projection layers to generate concept representations, which are then processed by an LLM to produce captions.



LLM Output Rep. → Image Output Projection

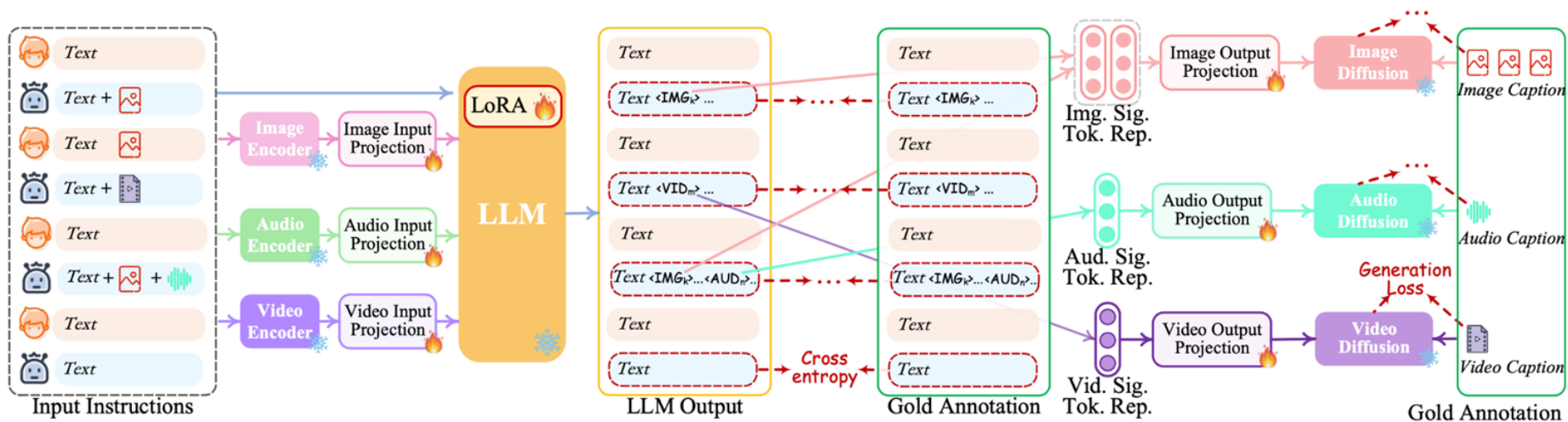
LLM Output Rep. → Image Output Projection



í Lw€3d -€İ€Ĉ/△△ⁿ Ę

≈š ħıı ŷš. □ □ □
| № š Ъ ÷ Ъќ □ ▯

Modality Switching Instruction Tuning (MosIT)





í Lw€3d ¬€İ Lū½ſ½ĩđ ¶Ë

Text + Video → Text + Image + Audio + Video



f ĩ, ↑½İĒĒ Ē ¶ĵ Ē½Ēđ ¶ĴİĒ
İ↑đĒĒđŚ ¶Ķ



İĒ↓ŚĒ ¶ĵ Ē½Ē ½Ēİ !İ↑đĒĒđŚ İĒĒ
İ↑½İĒĒ Ē ĒİĒĒ ¶ĵ Ē½Ē½İŚ ¶ĶŚ
pİĒĒ ¶İĒ ¶Ĵ ¶Ĵ ¶Ĵ İĒŚŚĒĒ Ē İ ¶ĵĒ
Ē½İŚ ¶ĶİĒĒ Ē ĒĒĶİĶŸ ½ĶĒ
¶İĒđ Śİ İĒŚİĒ½İŚ ¶ĶİĒ ĒĒ
Í ½YL ŚES

NEXT-GPT: Evaluation

€ŚŵĩĚĚ ģŦŚ Ę→ €ŚŵĩĚĚp || ½ ŚĚĚ Ĵ ņ d ĚĚ ģŦŚ



f ĩ k Ě ůĚá Ĵ ĴĚĚŋŋ Ę Ě || ŚĚ
|| Ĵ ĘĚĚ Ě || ½ Ĵ ŚĚĚ ĘĚĚĚŚ Ĵ ŚĚ
Ĵ ĴĚĚ || Ĵ ŚĚĴ Ĵ ĴĚĚ



Certainly! Here's a synthesized music track that could enhance the humor and lightheartedness of the scene:





Video Question Answering

Video Question Answering (VQA) is a task where a model is given a video clip and a question, and must generate an answer. This is a challenging task because the model must understand the visual content of the video and the context of the question.

Model	Version	Image Captioning			Image Question Answering			Comprehensive	
		NoCaps	Flickr 30K	COCO	VQA ^{v2}	VizWiz	OKVQA	MMB	SEED
InstructBLIP (Dai et al., 2023)	Vicuna-7B	<u>123.1</u>	82.4	102.2	-	33.4	33.9	36.0	-
LLaVA (Liu et al., 2023b)	LLaMA-2-7B-Chat	120.7	<u>82.7</u>	-	-	-	-	36.2	-
mPLUG-Owl (Ye et al., 2023b)	LLaMA-7B	117.0	80.3	<u>119.3</u>	-	39.0	-	46.6	<u>34.0</u>
Emu (Sun et al., 2023)	LLaMA-7B	-	-	117.7	40.0	35.4	34.7	-	-
DREAMLLM (Dong et al., 2023)	Vicuna-7B	-	-	115.4	56.6	45.8	44.3	49.9	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	-	-	74.7	<u>48.1</u>	-	60.9	-
NExT-GPT	Vicuna-7B	123.7	84.5	124.9	<u>66.7</u>	48.4	52.1	<u>58.0</u>	57.5

Model	Version	Video Captioning	Video Question Answering			Audio Captioning
		MSR-VTT	MSVD-QA	MSRVTT-QA	NExTQA	AudioCaps
Codi (Tang et al., 2023)	-	<u>74.4*</u>	-	-	-	<u>78.9*</u>
UIO-2XXL (Lu et al., 2023)	6.8B	48.8*	41.5	52.1	-	48.9*
Video-LLaMA (Zhang et al., 2023c)	LLaMA-7B	-	51.6	-	<u>29.6</u>	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	70.7	<u>59.2</u>	-	-
Emu (Sun et al., 2023)	LLaMA-7B	-	32.4	14.0	6.8	-
NExT-GPT	Vicuna-7B	76.2*	<u>64.5</u>	61.4	50.7	81.3*

◦ ō ŝ | - | ŝĩĩũ ūŝĉũŝũĕ

ĩĉŝŝũ½łĉ ħĩĉ ĵ n ĵĕ
ĩĵŝĕ ĉĕĩĩũ ħĩĵ ĵĩŝ



VideoPoet: Overview

ĤĈŠĤŪ½ĤĤ ĤĤĤĤ ĤĤ ĤĤ ĤĤ ĤĤ
ĤĤĤĤ ĤĤĤĤ ĤĤĤĤ ĤĤĤĤ ĤĤĤĤ

ĤĤ ĤĤ ĤĤĤĤ½ĤĤ ĤĤĤĤĤĤĤĤ
ĤĤĤĤ ĤĤĤĤ ĤĤĤĤĤĤĤĤ
ĤĤ ĤĤ

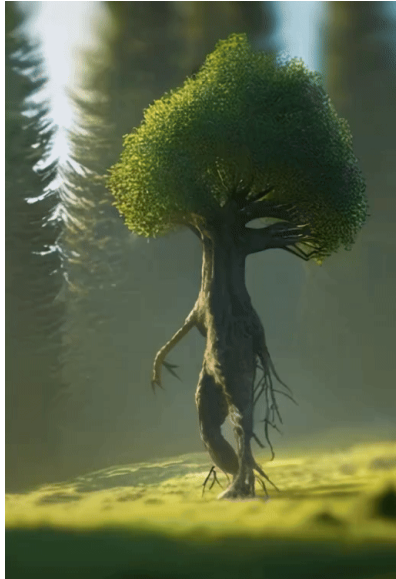


◦ Ń Ś | - | ŚĩĤũ ũŚĈũ ŚŨĚ

A tree walking through
the forest, tilt shift.

Ĥ | Ĵ ĚŚĚ ½ Ĥ Ĥ Ĥ ŚŚĚŚĚ
Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ
Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ

Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ
Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ
Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ Ĥ



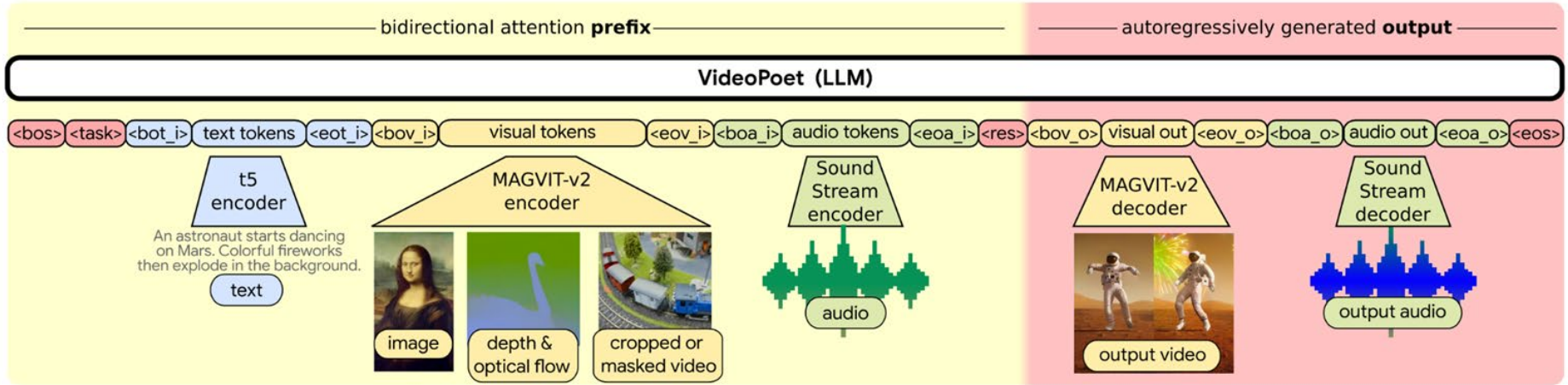


◦ VideoPoet (LLM)

• $\langle \text{bos} \rangle \langle \text{task} \rangle \langle \text{bot}_i \rangle$ text tokens $\langle \text{eot}_i \rangle \langle \text{bov}_i \rangle$

• language model backbone

• $\langle \text{eov}_i \rangle \langle \text{boa}_i \rangle$ audio tokens $\langle \text{eoa}_i \rangle \langle \text{res} \rangle \langle \text{bov}_o \rangle$ visual out $\langle \text{eov}_o \rangle \langle \text{boa}_o \rangle$ audio out $\langle \text{eoa}_o \rangle \langle \text{eos} \rangle$



◦ $\langle \text{bos} \rangle \langle \text{task} \rangle \langle \text{bot}_i \rangle$ text tokens $\langle \text{eot}_i \rangle \langle \text{bov}_i \rangle$ visual tokens $\langle \text{eov}_i \rangle \langle \text{boa}_i \rangle$ audio tokens $\langle \text{eoa}_i \rangle \langle \text{res} \rangle \langle \text{bov}_o \rangle$ visual out $\langle \text{eov}_o \rangle \langle \text{boa}_o \rangle$ audio out $\langle \text{eoa}_o \rangle \langle \text{eos} \rangle$

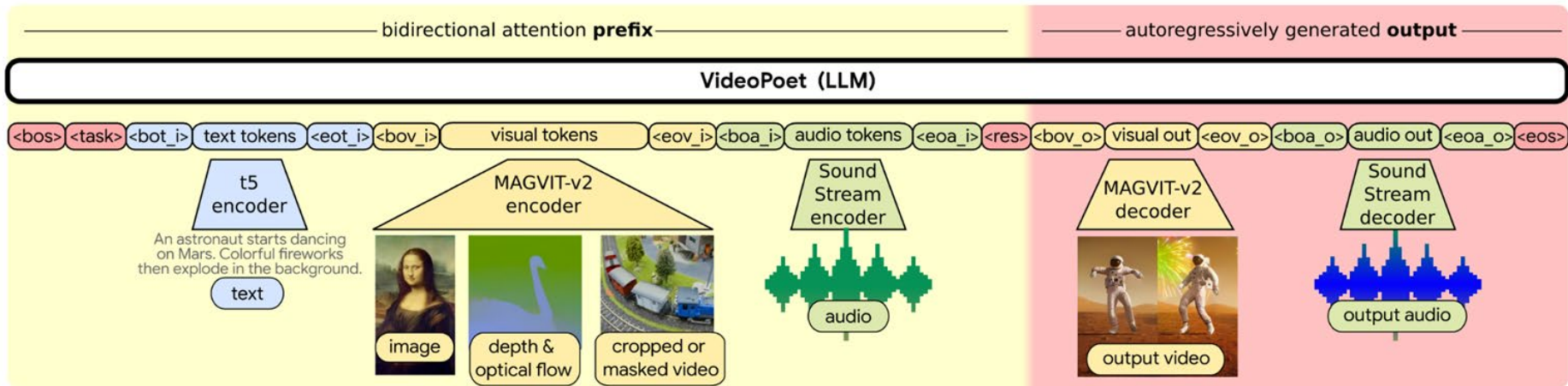
◦ $\langle \text{res} \rangle \langle \text{bov}_o \rangle$ visual out $\langle \text{eov}_o \rangle \langle \text{boa}_o \rangle$ audio out $\langle \text{eoa}_o \rangle \langle \text{eos} \rangle$

° Ń Š | - | Š Ĩ İ İ k □ ↓ Ë ≈ Š È | † İ 4 | =

How to decode visual features? $17 \times 896 \times 512 \longrightarrow 35,840$

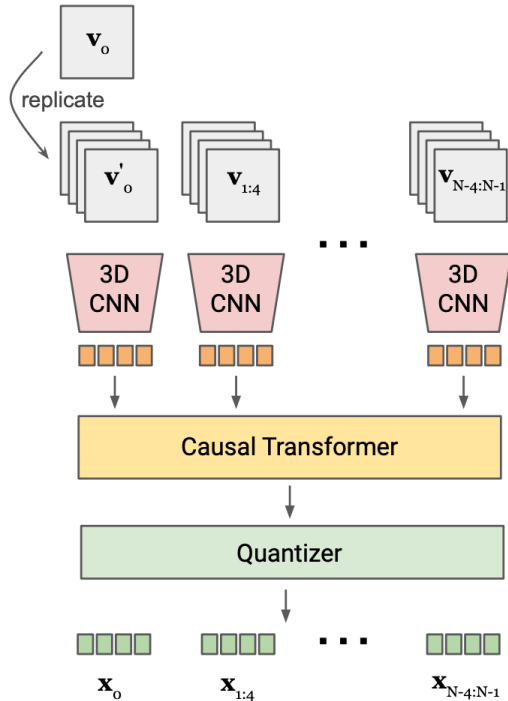
İ | π Š = È

impractical autoregressive sampling





◦ $\mathcal{S} \rightarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$



Casual tokenizer architecture comparison

The decoders, which are omitted from the figure, employ an architecture that is symmetric to the encoder

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

$\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ $\mathcal{S} \leftarrow \mathcal{S} \text{ i l l e } d \text{ } \circ p \in 3 \bar{u}$

◦ ǎ Š | - | Š ĩ ĩ ě ǎ ě ě ½ ě ½ ě ě / ě ě Š ě

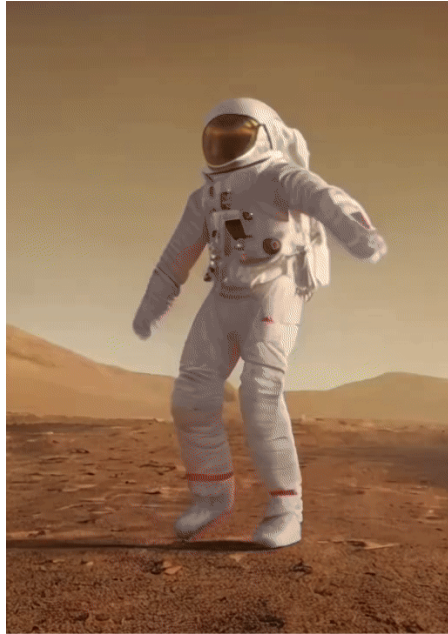


p ʒ ĩ ě ǎ Š |
 € ũ | ě ě ě | | ʒ ě | ʒ ě | | ě | ě ě ě ě | ʒ ě
 ½ ě | | ě ʒ ě ě ě | ½ ě ě ě ě ě | ě ʒ ě ě ě ě ě ě
 ◦ ě Š ě ě ě ě ě ě ě

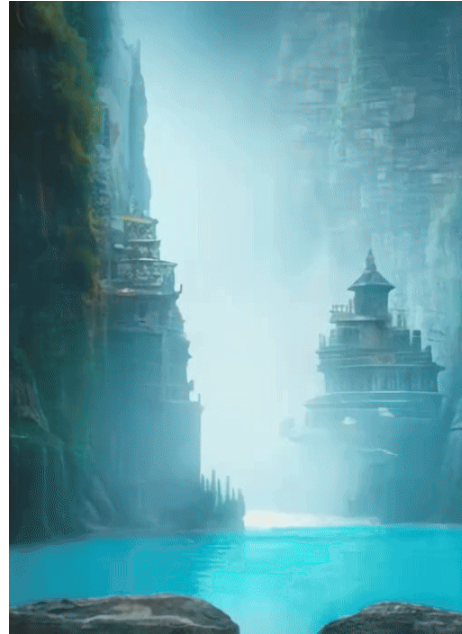
Extended Video

Two raccoons on motorbikes. A meteor shower falls behind the raccoons. **The meteors impact the earth and explode**

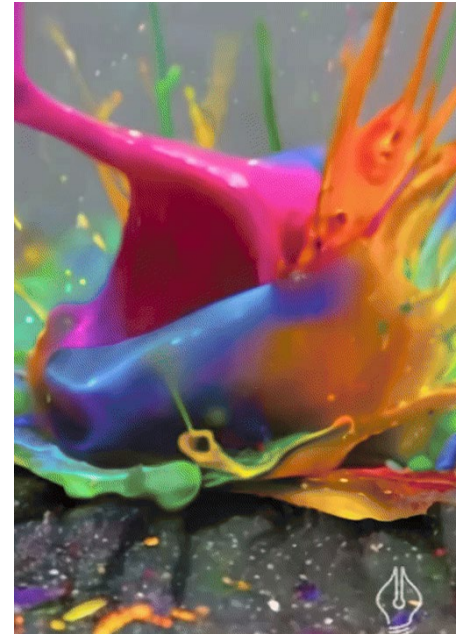
° ǎ Š | - | Šĩĩĩ | ʒ ǎ Š | ě Š ʒ Š Ć / ʒ | ʒ



ʒ ě ʒ | ʒ / ʒ ĩ ĩ ĩ / ʒ ě ě ʒ ʒ ʒ | ʒ ě
 ě ʒ Ć ě / ě ě | - Ć ů ʒ Ć ů | Ć ě ě
 Š w ʒ - ĩ Š ě ĩ ʒ ě ʒ ʒ | ʒ ʒ | ʒ ĩ Š



c - ° ě Ć ʒ Š ě | | ĩ ʒ / Š ĩ ! ě ě Š Ć ů ě ě ʒ / Ć ě ě
 Š ʒ Š - ě ʒ ĩ ů ! ě ě | ʒ Š ě ĩ ʒ Š ě ʒ ʒ ʒ Š ě ů ʒ ě ě
 ĩ Ć ʒ / ʒ ě | ʒ Š Ć ě Š Ć ů / ʒ Š Ć ʒ / ʒ ě ʒ ě ʒ ě Š ě
 ě ĩ Š Š ° ě Š Ć ů ʒ / ʒ ʒ ʒ ě ʒ / ʒ Š ě Š



ʒ Ć Š ě | - ĩ | ! ě Š w ʒ - ĩ ʒ ʒ ě
 ě ʒ ě ʒ ʒ ě ʒ ʒ | ů ě ʒ ʒ ě ů ʒ ě ʒ ě
 ʒ / ° ʒ ě | Š Ć ʒ ě ě ʒ

° ǎ Š | - | Šĩĩp || 1/2ⁿ Š3ĩ | 3° ǎ Š |

A geyser spraying water into the air.

ĩ | || 1/2ⁿ ĩ 1/2ⁿ ĩ S



